

计算社会学

数据时代的社会研究

SOCIAL RESEARCH IN THE DIGITAL AGE

荣获美国出版商协会
2019年
专业与学术杰出出版奖

[美] 马修·萨尔加尼克（Matthew J. Salganik）◎著
赵红梅 赵婷 ◎译

从观察、提问、实验
到开展大规模协作

为你提供一种
测量世界的
全新思维工具

站在数据科学和社会科学的十字路口
探察计算社会学的“美丽新世界”

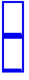
社会科学家的跨界指南之书
数据科学家的素养提升之作

萨尔加尼克是一位“纯然的计算社会学家”。

戴维·拉泽（David Lazer）
《科学》杂志

BIT BY BIT

中信出版集团



前言

这本书源于2005年在哥伦比亚大学一个地下室中发生的事情。那时，我还是一名研究生，正在为最终的毕业论文做一项在线实验。有关这项实验的学术部分我将在第4章进行介绍，但现在，我将告诉你们一件我的毕业论文或其他任何论文都未涉及的事情。这件事从根本上改变了我对研究的看法。一天早晨，当我来到位于地下室的工作室时，我发现一夜之间竟有约100个来自巴西的人参与了我的实验。这一简单的经历对我产生了深远的影响。当时，我的几个朋友正醉心于传统的实验室实验。我知道他们必须很费心地有偿召集并指导人们来参与实验，如果一天能有10个人完成实验，就算是不错的进展了。但对我的在线实验来说，我在睡觉的同时，就有100个人参与了实验。也许，一边睡觉一边做研究听起来美好得令人难以置信，但这是事实。技术的变化，尤其是技术从模拟时代到数字时代的转变，意味着我们可以用新的方式搜集和分析社会数据。这本书就是关于如何利用这些新方式开展社会研究的。

这本书是写给那些“想研究更多数据科学的社会科学家”和“想研究更多社会科学的数据科学家”以及对这两个领域的结合感兴趣的人的。因此，这本书的目标受众自然就不仅限于学生和教授了。尽管我目前在普林斯顿大学任职，但之前也在政府部门（美国人口调查局和技术产业领域的微软研究院）工作过，因此我知道，在大学之外同样存在着很多令人兴奋的研究。因此，只要你觉得自己正在做的是社会研究，那么无论在何处就职或使用何种技术，你都可以参阅此书。

你可能已经注意到了，这本书的语言风格和许多其他的学术著作稍有不同。这其实是我特意做的一个改变。这本书的部分内容源于我

从2007年起在普林斯顿大学的社会学系带领的一个“计算社会学”（Computational Social Science）研究生研讨班，因此我希望它能反映这个研讨班的一些活力和激情。具体而言，我希望这本书能够具备以下三个特点：有帮助的、面向未来的以及乐观的。

有帮助的：我的目标是写一本对你们有帮助的书。因此，我将以坦诚的态度、非正式的写作风格，通过实例阐述我的观点。我最想传达的是一种特定的思考社会研究的方式，而经验告诉我，传达这一思考方式的最好的方法就是采取非正式的写作风格并列举大量例子。此外，在本书的参考文献中，有一个部分叫“拓展阅读”，它旨在帮你过渡到有关我所介绍的多个主题的更加详细、更加专业的内容上。最后，我希望此书能对你们自己开展研究以及评估别人的研究有所帮助。

面向未来的：我希望这本书能帮助你利用现有的以及未来将出现的数字系统开展社会研究。我是从2004年开始做这类研究的，这期间数字系统发生了诸多变化，我坚信在你们的职业生涯中，你们也会感受到数字系统的许多变化。因此，要想让这本书“以不变应万变”，就要做到“抽象”。例如，这本书不会教你如何使用现有的推特应用程序界面（Twitter API），相反，它会教你如何受益于大数据资源（第2章）。这本书不会详细告诉你如何利用亚马逊土耳其机器人（Amazon Mechanical Turk，以下简称机器人MTurk）开展实验步骤，相反，它将教你如何设计和解读依赖于数字时代基础设施的实验（第4章）。通过采用这种抽象化的手法，我希望这本书能够成为一本主题适时、经得起时间考验的书。

乐观的：本书涉及两个群体——社会科学家和数据科学家，他们有着截然不同的背景和兴趣。除了书中将要介绍的科学方面的差异以外，我还发现，这两个群体看待事物的态度也是不同的。数据科学家一般而言是满怀希望的，而社会科学家一般而言是更具批判性的。也

就是说，同样是半杯水，数据科学家看到的是还有半杯水，而社会科学家看到的则是杯子有一半已经空了。在本书中，我将采取数据科学家的乐观态度。因此，在描述相关实例时，我将告诉你们在我看来这些例子的可取之处。当然，鉴于没有研究是完美的，我也会指出它们的问题所在，但我会尽力用乐观积极的方式指出。我不会为批判而批判，我的批判是为了能让你们设计出更好的研究。

我们仍处于数字时代社会研究的早期阶段，但我已经发现了一些普遍存在的误解，它们的普遍程度让我觉得有必要在前言中对其进行说明。就数据科学家而言，我发现他们有两个常见的误解。第一个是认为数据越多越有利于解决问题。但对社会研究来说，我的经验告诉我并不是这样的。事实上，对社会研究来说，好的数据似乎要比更多的数据更有帮助。第二个是数据科学家通常认为社会科学只不过是一堆围绕常识的花言巧语罢了。当然，作为一名社会科学家，更确切地说是社会学家，我不同意这样的观点。聪明的人长期以来一直在努力理解人类的行为，因此忽视这一努力所取得的成果似乎是不明智的。我希望通过这本书，以一种易于理解的方式和你们分享其中的一些成果。

就社会科学家而言，我发现他们也有两个常见的误解。第一个是有些社会科学家会因为少数不真实的数据而彻底否定使用数字时代的工具开展社会研究这一观念。如果你正在读这本书，那你可能已经读过许多平庸地或错误地（或两种方式都有）使用社交媒体数据的论文。我也读过。但是如果因为这些论文就得出结论，说数字时代的社会研究都是不好的，这将是一个严重的错误。事实上，你可能也读过许多平庸地或错误地使用调查数据的论文，但你并没有因此而否定所有使用调查数据的论文。这是因为你知道，也有使用调查数据并且做得很不错的研究。而我将通过这本书告诉你们，使用数字时代的工具并且做得很不错的研究也是有的。

我所发现的社会科学家的第二个常见误解是容易将现在和未来混淆。当我们对数字时代的社会研究，即我在本书中将探讨的研究，进行评估时，思考以下两个截然不同的问题至关重要：“这类研究现在做得怎么样”以及“这类研究将来会做得怎么样”。研究人员会被训练来回答第一个问题，但对这本书而言，我认为更重要的是第二个问题。也就是说，尽管数字时代的社会研究尚未做出巨大的、改变范式的贡献，但数字时代社会研究的进步速度快得惊人。因此，相比于其目前的发展水平，它的变化速度更让我感到兴奋不已。

尽管上一段似乎是在告诉你们，数字时代的社会研究可能会在未来的某个时间变得相当成功，但我的目标并不是向你们推销任何特定类型的研究。我个人并未持有推特（Twitter）、脸谱网（Facebook）、谷歌（Google）、微软（Microsoft）、苹果（Apple）或其他任何科技公司的股份。但是，为了做到充分披露，我应该告诉你们我曾在微软、谷歌和脸谱网工作过或是接受过其研究经费赞助。因此，在整本书中，我的目标是让自己做一个可信的叙述者，告诉你们所有可能的令人兴奋不已的新事物，同时引导你们避开一些我曾看到有人掉进去的陷阱（有的我自己也曾掉进去过）。

社会科学和数据科学的交叉学科有时会被称为“计算社会学”。有些人认为这是一个技术领域，但这本书并不是传统意义上的技术图书。例如，这本书的正文中并没有公式。之所以选择这样的方式，是因为我想呈现对数字时代社会研究的一个全面的看法，其中包括大数据资源、调查、实验、大规模协作和道德伦理。但事实证明，涵盖所有这些主题并提供每个主题中详细的技术细节是不可能的。相反，我会在本书参考文献中的“拓展阅读”里推荐更多的技术资料。换句话说，这本书不是为了教你如何做某种特定的计算，而是为了改变你对社会研究的思考方式而写的。

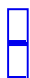
如何在教学中使用这本书？正如前面所述，本书的部分内容来自我从2007年开始在普林斯顿大学带领的一个“计算社会学”研究生研讨班。你们可能想用这本书进行教学，所以我觉得有必要解释一下我是如何将源于课堂的素材写成这本书的，以及我想象的这本书在其他课堂中的使用方式。

有几年时间，我上课是没有指定教材的，我只是给学生指定一些文章。虽然他们能够从这些文章中学到东西，但只学习这些文章还不足以让他们发生我所期待的观念转变。所以我会用课堂大部分的时间讲述这些文章的背景，讲述应该采取怎样的视角以及给予他们建议，进而帮助学生获得更全面的认识。在这本书中，我试图以不涉及社会科学或数据科学专业知识的方式记录上述所有的背景、视角和建议。

对于为期一学期的课程，我建议将这本书与其他各种阅读材料配套使用。例如，课程可能会花两周时间来做实验，这时你可以使用第4章的内容，同时选取诸如以下主题的阅读材料：预处理信息在实验设计和分析中的作用；在公司大规模的A/B测试过程中所浮现出来的统计和计算问题；实验设计，尤其是原理方面，以及与通过机器人MTurk这样的在线劳动力市场招募实验参与者相关的实践、科学和伦理方面的问题。你也可结合编程方面的阅读材料或活动。至于如何从这些材料中选出合适的配套材料，就取决于你的学生（是本科、研究生还是博士）以及他们的背景和目标。

在一个为期一学期的课程中，你也可以每周给学生分配一些任务。这本书的每一章都会涉及各种各样的“活动”，我将把“活动”放在参考文献中，同时我也标注了它们的难度等级：简单（）、中等（）、困难（）以及非常困难（）。此外，我还标注了每个问题所需的技能：数学（）、编码（）以及数据采集（）。最后，对一些我个人比较喜欢的活动，我会备注心形图标（）。我希望在这么多的任务活动中，你能找到适合自己的。

为了帮助人们在教学中使用这本书，我已经开始搜集相关的教学资料了，例如教学大纲、幻灯片、每章推荐的配合材料以及一些任务活动的解决方案。你可以访问http://www.bitbybitbook.com查看或完善这些资料。



第1章　简介

1.1 一处墨迹

2009年夏天，手机铃声响遍了整个卢旺达。除了来自家人、朋友和商业伙伴的数百万个电话之外，大约有1000名卢旺达人还接到了由乔舒亚·布卢门斯托克（Joshua Blumenstock）及其同事打来的电话。研究人员从卢旺达最大手机供应商的数据库中随机抽样进行调查，以完成对财富与贫困的研究，这个数据库中有150万名客户。布卢门斯托克和他的同事会询问这些被随机选中的人是否愿意参与调查，然后向其解释这项研究的性质，接下来便会询问一系列有关他们的人口学特征、社会特征和经济特征方面的问题。

到目前为止，我所描述的一切都让这项研究听起来像是一项传统的社会科学调查。但接下来我要描述的就不再传统了，至少目前来说是这样的。除了调查而来的数据外，布卢门斯托克和同事还拥有这150万人的完整通话记录。他们将这两部分数据结合起来，利用调查数据训练了一个机器学习模型，使模型能根据一个人的通话记录预测其财富状况。接着，他们利用这个模型评估数据库中150万名客户的财富状况，还利用通话记录中包含的地理信息判断这150万名客户的居住位置。最后他们将所有这些信息——估算的财富状况以及居住位置，综合到一起，绘制出高分辨率的卢旺达财富地理分布图。尤其是，他们能够估算出卢旺达2148个街区（该国的最小行政单位）中每一个街区的财富状况。

要证实这些估算是不可能的，因为从来没有人估算过卢旺达中如此小的地理区域的财富状况。但在布卢门斯托克和同事把这些估算值汇总为分别反映卢旺达30个地区财富状况的数值后，他们发现，这些数值与通过人口统计和健康调查（Demographic and Health Survey）

得到的数据非常接近，而人口统计和健康调查被认为是发展中国家调查的黄金标准。虽然这两种方法在此案例中产生了类似的结果，但布卢门斯托克和同事的方法要比传统的人口统计和健康调查的方法快了差不多10倍，成本为后者的1/50左右。这些明显更快、更节省成本的预测为研究人员、政府和公司创造了新的可能性（Blumenstock, Cadamuro, and On 2015）。

这项研究有点像一个罗夏墨迹测验[注]：人们看到的事物取决于他们的背景。许多社会科学家从中看到了一个新的测量工具，这个工具可以检验经济发展理论。许多数据科学家从中看到了一个很酷的、新的机器学习问题。许多商界人士看到了一个可以让他们从已经搜集到的大数据中获利的好方法。许多隐私权倡导者从中看到了一个可怕的警示：我们也许生活在一个大规模监控的时代。最后，许多政策制定者从中看到了新技术能够帮助我们创造一个更好的世界。其实，这项研究与这些都相关，而且正是因为它融合了这么多特征，所以我把它看作了解社会研究之未来的一扇窗。

---

1. 罗夏墨迹测验是一种著名的人格测验，它会向被试呈现由墨迹偶然形成的图案，让被试观看并说出由此联想到的事，研究人员由此对反应符号进行分析，从而判断被试的人格特征。——编者注

1.2 欢迎来到数字时代

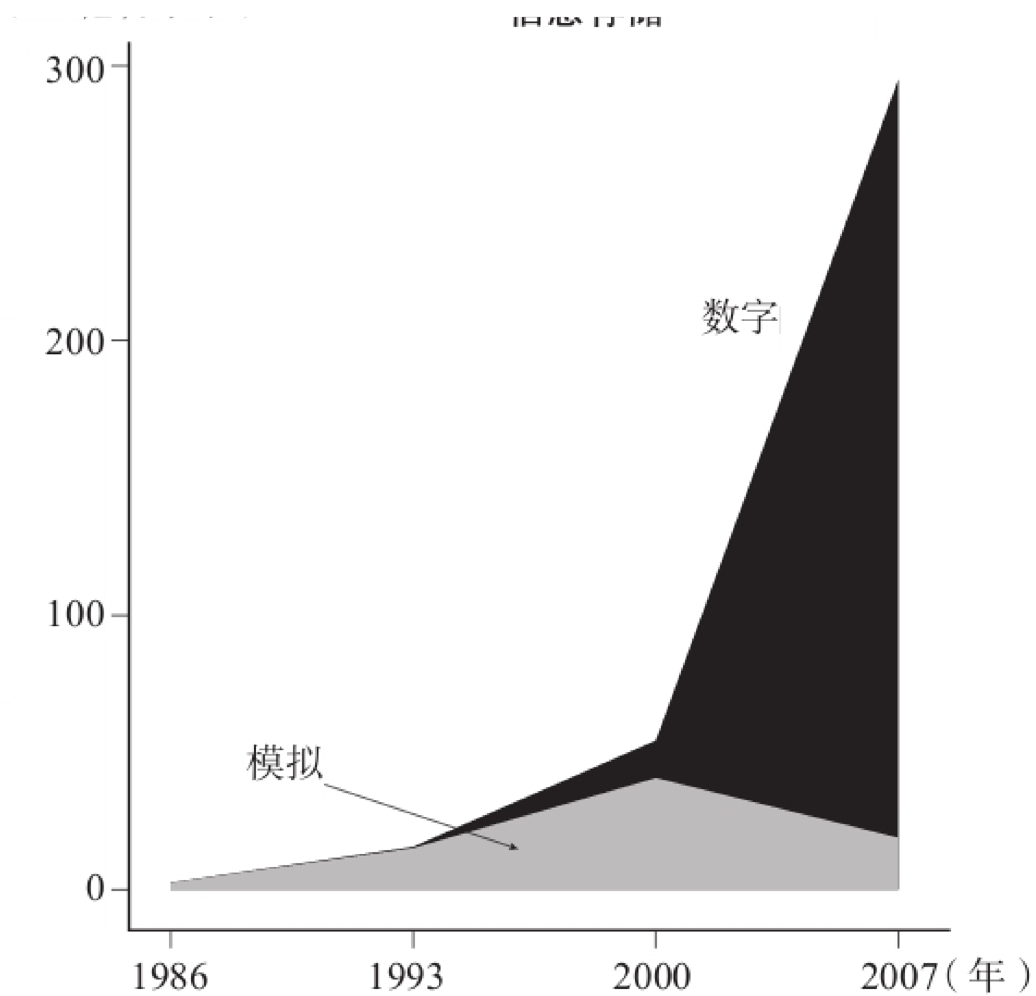
数字时代无处不在，它在不断发展，并且改变着研究的可能性。

这本书的核心前提是数字时代能为社会研究创造新的机会。研究人员现在能以不久前还几乎不可能的方式观察行为、提出问题、开展实验以及彼此协作。但新的风险也随之而来：研究人员现在能以过去绝不可能的方式去伤害人们。这些机会和风险源于从模拟时代到数字时代的转变。这种转变并不是像开灯那样瞬间就发生了，事实上，这种转变目前还尚未彻底完成。但目前为止发生的事情，已经足以让我们相信有大事正在发生了。

注意到这种转变的一个方法是观察发现你们日常生活中的变化。生活中，许多曾经是模拟的东西现在变成智能的了。也许你曾经用的是带胶卷的相机，但现在用的是数码相机（可能你们的智能手机就有数码相机的功能）。也许你们曾经读的是纸质的报纸，现在却在线看新闻。也许你们曾经用现金来付款，现在却是用信用卡。在上述每一种情况下，从模拟到数字的转变，都意味着更多关于你的信息被以数字化的形式获取并存储了下来。

事实上，总体来看，从模拟到数字的转变所产生的影响是非常惊人的。信息量正在迅速增加，更多的信息以数字化的形式被存储，进而便于分析、传输和归并。这些数字信息被称为“大数据”。在数字数据爆炸式增长的同时，有条件使用计算机的人的数量也在不断增加（图1.1）。这些趋势，即越来越多的数字数据以及越来越多的使用计算机的人，在可预见的未来很可能会持续下去。

（10亿吉字节）　　　　信息存储


300
200
100
0
数字
模拟
1986
1993
2000
2007（年）


（10 亿条，每秒）
计算能力
7 500
5 000
2 500

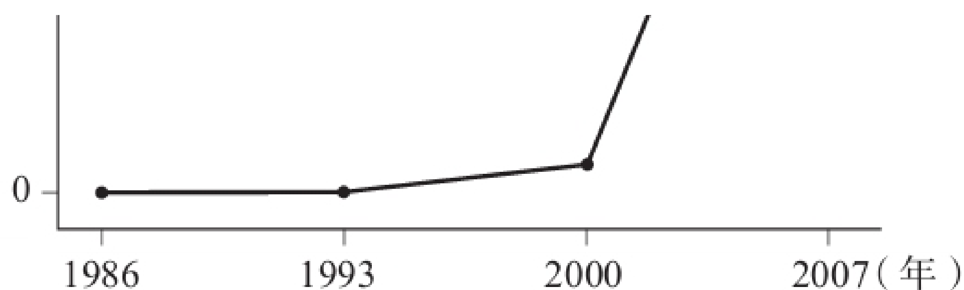


图1.1 信息存储能力和计算能力正在显著提高。此外，信息存储现在几乎已经全部数字化了。这些变化为社会研究人员创造了不可思议的机会。改编自Hilbert and López（2011）。

考虑到社会研究的目的，我认为数字时代最重要的特征就是计算机随处可见。从最初房间般大的、只有政府和大公司才能使用的计算机发展而来，今天的计算机，其尺寸在不断缩小，普及程度在不断增加。从20世纪80年代开始，每10年就会有一种新型的计算机诞生：个人计算机、笔记本电脑、智能手机以及现在“物联网”中的嵌入式处理器（即汽车、手表和恒温器等设备内部的嵌入式计算机）（Waldrop 2016）。除了计算，这些随处可见的计算机还开始具备越来越多的功能：感知、存储和传输信息。

对研究人员来说，利用网络最容易看出随处可见的计算机所产生的影响。互联网是一个被全面监测的环境，非常适合研究人员开展实验。例如，一个网上商店很容易就可以搜集到精确的数百万顾客的购买行为数据。然后，它可以随机选择几组顾客并为其提供不同的购物体验。这种在精准掌握了顾客购物数据的基础上再进行随机选择的能力，意味着网上商店能够持续开展随机对照实验。事实上，只要曾在网上买过东西，你的购买行为就已经被记录下来了，之后你也几乎必然地会成为某项实验的参与者，无论你自己是否知道。

这种被全面监测、完全随机化的情况不仅局限于网上，这在线下也开始变得越来越普遍。实体店已经搜集了非常详细的购买行为数据，同时它们也正在开发相关基础设施，以便追踪顾客的购买行为，并将实验研究结果用于日常商业活动中。“物联网”意味着现实世界

中的行为会越来越多地被数字传感器捕获。换句话说就是，当你思考数字时代的社会研究时，你不应该只想到“在线”社会研究，你应该想到它可以发生在任何地方。

数字时代使行为测量和实施随机化的处理成为可能，同时还为人们创造了新的交流途径。这些新的交流途径使研究人员能够开展创新性的调查，并与同事以及普通大众进行大规模协作。

怀疑论者可能会说，上述这些功能并不能算是真正意义上的新事物。也就是说，在过去，人们的交流途径也曾有过其他一些重大的进步，例如电报（Gleick 2011），而且自20世纪60年代以来，计算机的运行速度也基本上一直在以同样的速度增长（Waldrop 2016）。但这些怀疑论者所忽略的是，在某种程度上，多个相同的东西合起来会变成一个不同的东西（Halevy, Norvig, and Pereira 2009）。让我用我喜欢的一个类比来阐明这一观点：如果你能捕捉到一匹马在某一刻的画面，你便拥有了一张照片；而如果你在一秒内捕捉到一匹马的24个画面，那么你便拥有了一部电影的片段。当然，一部电影其实就是许多张照片，但只有顽固的怀疑论者才会坚持声称照片和电影完全一样。

研究人员正在进行一项类似于从摄影到拍电影的转变，但这一转变并不代表我们过去所学的所有东西都应该被摒弃。正如摄影的原理会影响拍电影的原理一样，那些在过去100年里发展起来的社会研究理论也会对未来100年的社会研究产生影响。这一转变意味着我们不应该一直做同样的事情。相反，我们必须把过去的、现在的以及未来的方法结合起来。例如，乔舒亚·布卢门斯托克和同事所做的研究就结合了传统的调查研究和一些人可能称之为数据科学的方法。单凭调查研究或是单凭通话记录都无法绘制出高分辨率的反映财富估值的地图，这两个是缺一不可的。更概括地说，社会研究人员需要将社会科学和

数据科学的思想结合起来，才能充分利用数字时代带来的机会，只靠其一是不够的。

1.3 研究设计

研究设计是将问题和答案联系起来。

本书是为两个群体所写的，这两个群体有很多需要互相学习的地方。一方面，这本书是写给社会科学家的，他们接受过社会行为研究方面的训练，也有这方面的经验，但对数字时代所带来的机会不是很熟悉。另一方面，这本书是写给那些能得心应手地使用数字时代工具，但对社会行为研究来说是新手的研究人员的。这类研究人员不愿被冠以一个简单的称谓，但我将把他们称为数据科学家。这些数据科学家经常会接受计算机科学、统计学、信息科学、工程学和物理学等方面的训练，已成为最早开展数字时代社会研究的一群人，这部分是因为他们能够获得必要的数据，同时也具备相应的计算能力。本书试图让这两个群体彼此合作，进而创造出比单独一个群体所能创造的更加丰富、更加有趣的东西。

要实现这一强有力的合作，最好的方式不是专注于抽象的社会理论或是花哨的机器学习。最好的起点是研究设计。如果你将社会研究看作询问和回答有关人类行为问题的过程，那么研究设计就是“结缔组织”，它能将问题和答案联系起来。而建立正确的联系是设计出令人信服的研究的关键。本书将重点介绍4种方法：观察行为、提问、开展实验以及与他人合作。这些方法你之前应该已经见过或可能用过，但特别之处在于，数字时代为我们带来了新的搜集和分析数据的机会。这些新机会要求我们将这些经典的方法现代化，但不是要取代这些方法。

1.4 本书的主题

本书的两个主题分别是：（1）将现成品和定制物结合起来；（2）道德伦理。

这两个主题将贯穿整本书，我之所以在这里强调它们，是为了让你们在其反复出现时能够注意到。第一个主题可以通过对比马塞尔·杜尚（Marcel Duchamp）和米开朗琪罗（Michelangelo）这两位伟人来阐述。杜尚主要是因为他的现成品艺术作品（例如《泉》）而闻名，这些艺术作品都是普通物品经他稍做修改而创造出来的。而米开朗琪罗则不是通过修改现成品进行创作的。当他想创作一尊大卫的雕像时，他并没有去寻找一块看起来像大卫的大理石，而是花了三年的时间雕刻出了他的杰作。因此，《大卫》不是一个现成品艺术作品，而是一个非现成品艺术作品（图1.2）。

这两种风格——现成品艺术作品和非现成品艺术作品，大致可以映射出数字时代的社会研究所采用的风格。正如你们将要看到的，本书中的一些例子就涉及对某些大数据资源的巧妙的重新配置，而这些大数据资源最初是由公司或政府所创建的。在其他例子中，研究人员则从一个特定的问题出发，然后使用数字化工具创建出回答该问题所需的数据。如果做得好的话，这两种模式都非常强大。因此，数字时代的社会研究将既包括现成品作品又包括非现成品作品，既包括杜尚又包括米开朗琪罗。

如果你们通常使用的是现成数据，那么我希望这本书能告诉你们非现成数据的价值。同样，如果你们通常使用的是非现成数据，那么我希望这本书能告诉你们现成数据的价值。最后，也是最重要的，我

希望这本书能告诉你将这两种数据结合起来使用的价值。例如，乔舒亚·布卢门斯托克及其同事就是杜尚和米开朗琪罗的结合体：他们把通话内容录音（一个现成数据）用于不同的用途，同时又创建了自己的调查数据（一个非现成数据）。在整本书中，你们都将看到现成品与非现成品的结合，这种结合往往既需要社会科学的思想也需要数据科学的思想，并且这种结合常常会带来最令人兴奋的研究。





现成品艺术作品　　　　非现成品艺术作品

图1.2 马塞尔·杜尚的《泉》和米开朗琪罗的《大卫》。《泉》就是一件现成品艺术作品，这类作品是艺术家对现实世界中已经存在的东西进行创造性加工修改后而产生的艺术品。而《大卫》则是有意创造的艺术品，是一件非现成品艺术作品。数字时代的社会研究将既包括现成品作品又包括非现成品作品。《泉》由艾尔弗雷德·施蒂格利茨（Alfred Stieglitz）摄于1917年（来源：The Blind Man, no.2/Wikimedia Commons）。《大卫》由约尔格·比特纳·翁纳（Jörg Bittner Unna）摄于2008年（来源：Galleria dell’Accademia, Florence/Wikimedia Commons）。

贯穿本书的第二个主题是道德伦理。我将告诉你们，研究人员如何利用数字时代的机会开展令人兴奋且意义重大的实验。同时我也将告诉你们，利用这些机会的研究人员将如何做出艰难的伦理决策。本

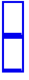
书第6章全部是关于道德伦理的，但其他章节也会涉及这一话题，因为在数字时代，道德伦理将成为研究设计中越来越重要的一个部分。

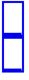
布卢门斯托克及同事的实验可以再次被用来证明这一点。150万人的通话记录为他们的研究创造了很好的机会，但同时也提供了造成伤害的机会。例如，乔纳森·迈耶（Jonathan Mayer）及同事在2016年已经表明，即使是对“匿名化”的通话内容录音（即没有名字和地址的数据），在结合公开信息后，研究人员也可能从中识别出属于某些特定人员的通话内容，进而推断出有关他们的敏感信息，例如某些健康状况的信息。也就是说，虽然布卢门斯托克及同事并未试图找出某些特定的人并推断有关他们的敏感信息，但这一可能性会让他们很难拿到通话数据，这迫使他们在进行研究时要采取全面的保护措施。

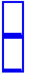
除了详细的通话记录以外，数字时代的许多社会研究中都存在一个让人很不安的根本性问题：研究人员（经常与公司和政府合作）对实验参与者的生活拥有越来越强的控制力。我所说的控制力是指在未征得参与者同意，甚至在他们不知情的情况下，对他们做一些事情的能力。例如，研究人员现在可以观察数百万人的行为，而且正如我后文将描述的，研究人员也可以让数百万人参加大规模的实验。所有这些事情都可能在未征得当事人同意或其不知情的情况下进行。研究人员的控制力在不断增强，对如何使用这一控制力的规范却没有相应地变得更加明确。也就是说，研究人员必须在彼此不一致、相互重叠的法律法规的基础上决定他们该如何行使这一控制力。因此，即便是善意的研究人员，在面对强大的控制力和模糊的指导方针时，可能也会被迫去做一些艰难的抉择。

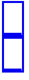
如果你们通常关注的是数字时代的社会研究所创造的新机会，那么我希望这本书能让你们明白这些机会也会带来新的风险。同样，如果你们通常关注的是这些风险，那么我希望这本书能帮助你们发现新机会（需要冒险的机会）。最后，同时也是最重要的，我希望这本书

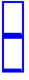
能帮助大家负责任地平衡数字时代的社会研究所带来的机会和风险。当研究人员开始拥有更强的控制力时，他们也必须承担更大的社会责任。

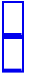












2.5 结论

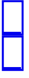
大数据资源无处不在，但利用它们进行社会研究可能会遇到很多问题。根据我的经验，“天上不会掉馅饼”这类规则也适用于社会研究的数据：如果你不投入大量工作来搜集数据，那么你可能需要投入大量工作去思考和分析数据。

今天的大数据资源往往具有以下10个特征，未来的大数据资源也可能如此。其中有3个特征通常是（但并不总是）有助于研究的：海量性、持续性以及不反应性。而其余7个则通常是（但并不总是）不利于研究的：不完整性、难以获取、不具代表性、漂移、算法干扰、脏数据以及敏感性。其中许多特征之所以会出现，是因为大数据资源并不是为了社会研究而创建的。

基于本章的观点，我认为以下三点最能说明大数据资源在社会研究中的价值。首先，大数据资源能让研究人员验证两个互相矛盾的理论预测的正误，比如法伯的纽约市出租车司机研究。其次，基于大数据资源的临近预测能为决策者提供更好的评估信息，该类研究的一个事例是金斯伯格等人的谷歌流感趋势研究。最后，大数据资源有助于研究人员在不开展实验的情况下进行因果推断，该类研究的事例包括马斯和莫雷蒂针对同侪效应对生产力影响的研究以及埃纳维等人针对易贝上起拍价对拍卖影响的研究。然而，无论是上述哪一种情形，都需要研究人员赋予数据更多东西，例如确定对因果推断很重要的研究定量或两个观点互相矛盾的理论。因此，我认为对大数据资源的作用的最好描述是，它们能够帮助那些可以提出有趣且重要问题的研究人员。

本章结束之前，我认为还有一个问题值得思考，那就是大数据资源可能对数据和理论之间的关系产生重要的影响。目前为止，本章所采用的方法是理论导向的实证研究。但大数据资源也能让研究人员进行实证导向的理论推理。也就是说，通过仔细积累经验事实、实际模式和难解之题，研究人员可以建立新的理论。这一非传统的、在没有理论假设的情况下直接从数据入手建立理论的方法并非是首次被提及，巴尼·格拉泽（Barney Glaser）和安塞尔姆·施特劳斯（Anselm Strauss）在其提倡扎根理论的著作中就对这一方法进行了最有力的阐述。但这种直接从数据入手的方法并没有像数字时代的一些有关研究的文章所宣称的那样意味着“理论的终结”（Anderson 2008）。相反，随着数据环境的变化，我们应该期望数据和理论的关系能重新得到平衡。在过去，数据采集是很昂贵的，因此只搜集那些理论表明最有用的数据是说得通的。但现在，我们有海量的可以免费使用的数据，因此除了搜集最有用的数据以外，尝试直接从现有数据入手也是可以的（Goldberg 2015）。

本章内容表明，研究人员通过观察人类可以了解到很多东西。在接下来的几章中，我将介绍如何通过调整数据采集的方法，通过提问（第3章）、开展实验（第4章），甚至直接选择一部分人参与研究过程（第5章）这些与人们更直接的互动了解更多不同的东西。



第3章　提问

3.1 简介

因为不能问海豚问题，所以关注海豚的研究人员不得不通过观察其行为的方式来了解这一物种。而研究人类则相对容易一些，因为人类会说话。与人交谈在过去是社会研究的一个重要组成部分，我希望将来也是如此。

在社会研究中，与人交谈通常有两种形式：调查和深度访谈。简单来说，通过调查进行的研究需要系统地招募大量参与者，需要高度结构化的调查问卷以及使用统计方法实现从参与者到更大群体的泛化。而通过深度访谈进行的研究则通常需要少量的参与者和半结构化的对话，最终得出关于参与者的丰富的定性描述。调查和深度访谈都是很有效的方法，但从模拟时代到数字时代的转变对调查的影响更大。因此，在本章中，我将重点介绍调查研究。

本章将向大家展示，数字时代为调查研究人员创造了许多令人兴奋的机会，使他们能以更快的速度、更低的成本搜集数据，能提出不同类型的问题，并能利用大数据资源放大调查数据的价值。技术变革可以改变调查研究已经不是新鲜事了。大约在1970年，电话这一新通信技术的诞生也引发了一场类似的变革。幸运的是，理解电话如何改变调查研究有助于我们预测数字时代将如何改变调查研究。

今天我们所认可的调查研究起源于20世纪30年代。在调查研究的第一个时代，研究人员会随机选取地理区域（如城市街区），然后前往这些地区，与随机选取的住户进行面对面的交谈。之后，固定电话开始在一些富裕国家大量涌现，这一技术发展最终促使调查研究进入了第二个时代。在该时代，无论是人们被选为样本的方式还是对话发

生的方式都发生了改变。研究人员不再选取某地理区域的住户作为样本，而是通过一个随机拨号的程序随机选取电话号码作为样本。他们也不再前往当地和人们面对面交谈了，而是通过打电话来交流。这些看似很小的组织实施上的变化却让调查研究变得更快、成本更低、更加灵活。除了这些益处之外，这些变化也引发了争议，因为许多研究人员担心这些取样和访谈方式的改变会导致各种偏差。但最终，在大量工作之后，研究人员找到了通过随机拨号和电话访谈搜集可靠数据的方法。因此，成功借助了社会上的技术基础设施，研究人员能以现代化的方式进行调查研究了。

现在，另一项技术发展——数字技术，最终将带领我们进入调查研究的第三个时代。这一转变的部分原因是第二个时代的方法逐渐不再适用了（Meyer, Mok, and Sullivan 2015）。例如，由于各种技术和社会原因，多年来无回答率（即样本中未参与调查的受访者的比例）一直在上升（National Research Council 2013）。这一长期趋势意味着，如果现在开展电话调查，那么无回答率会超过90%。

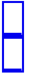
另一方面，向第三个时代的过渡在部分程度上也受到了令人兴奋的新机会的推动，我将在本章对其中一些机会进行描述。尽管目前还没有定论，但我预计非概率抽样、计算机管理的调查以及使用大数据资源进行调查，将成为调查研究第三个时代的特征（表3.1）。

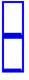
表3.1 调查研究的三个时代

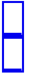
	抽样	访谈	数据环境
第一个时代	区域概率抽样	面对面调查	单独调查
第二个时代	随机拨号概率抽样	电话调查	单独调查
第三个时代	非概率抽样	计算机管理的调查	使用大数据资源进行调查

调查研究第二个时代和第三个时代之间的过渡并不是一帆风顺的，关于研究人员应如何继续开展调查研究，一直存在激烈的争论。回顾第一个时代和第二个时代之间的过渡，我认为对今天的我们来说，很关键的一个经验是：开始并不是结束。也就是说，在第二个时代，许多基于电话的方法起初都是临时决定的，而且也不太有效。但经过努力，研究人员解决了这些问题。例如，在起初的许多年里，研究人员一直在摸索随机拨号，然后才产生了沃伦·米托夫斯基（Warren Mitofsky）和约瑟夫·韦克斯伯格（Joseph Waksberg）具有良好实用性和理论基础的随机拨号抽样法（Waksberg 1978; Brick and Tucker 2007）。因此，我们不应该认为第三个时代的方法在当前的状态就是其最终状态。

调查研究的发展历程表明，这一领域的发展是由技术和社会的变化所驱动的，我们无法阻止这一发展。我们应该欣然接受这一发展，并继续从之前的时代汲取智慧，这也是我在本章将遵循的理念。首先，我认为大数据资源不会取代调查，而且其丰富性还会提升而不是降低调查的价值（3.2节）。鉴于这一点，我将总结在调查研究的前两个时代发展起来的调查误差总框架（3.3节）。该框架能让我们了解有关代表性的新方法，尤其是非概率样本（3.4节）以及测量的新方法，特别是向受访者提问的新方法（3.5节）。最后，我将描述两个将调查数据和大数据资源结合起来的研究模板（3.6节）。







3.5 提问的新方法

传统的调查是不公开的、无聊的，并且远离生活。而如今，我们可以问一些更开放、更有趣、更贴近生活的问题。

调查误差总框架促使研究人员将调查研究作为一个由两部分组成的过程来思考，这两个部分分别是招募受访者和向他们提问。在3.4节中，我介绍了数字时代是如何改变我们招募受访者的方式的，而现在我将介绍数字时代如何让研究人员能以新的方法来提问。这些新方法可以被用于概率样本或非概率样本。

调查方式是关于问题传递的方式或渠道的，它对测量有重要的影响（Couper 2011）。在调查研究的第一个时代，最常见的方式是面对面，而在第二个时代，则是电话。有些研究人员将调查研究的第三个时代看作仅仅是调查方式的增加，新增了计算机和手机。然而数字时代不仅仅意味着问题和答案传递渠道的改变，从模拟到数字的转变使研究人员能够改变其提问的方式。

迈克尔·肖伯（Michael Schober）和同事的一项研究可以说明调整传统方法以使其更好地匹配数字时代通信系统的好处。在这项研究中，肖伯和同事比较了几种不同的利用手机向受访者提问的方法。其中一种是语音会话，该方法可以说是调查研究第二个时代方法的延伸；另一种是通过短信发送的微观调查，该方法没有什么广为人知的先例。然后他们发现，相比于语音会话，通过短信发送的微观调查能搜集到质量更高的数据。换句话说，只是简单地用新媒介来实施旧方法，是无法搜集到最高质量的数据的。相反，通过清楚地思考手机相

关的功能和社会规范，肖伯和同事找到了一种更好的提问方式，进而搜集到了更高质量的答案。

研究人员可以从多个维度对调查方式进行分类，但我认为数字时代调查方式最主要的特征是通过计算机管理，而不是由采访者管理（例如电话和面对面访谈调查）。采访者不参与数据采集过程有诸多好处，这同时带来了一些挑战。就好处而言，采访者不参与数据采集可以减少社会期望偏差，而社会期望偏差会使受访者倾向于以最好的方式来呈现自己，例如谎称自己没有做过滥用药物等被社会污名化的行为，或谎称自己做过投票等被提倡的行为（Kreuter, Presser, and Tourangeau 2008）。采访者不参与数据采集还能消除采访者的影响，即采访者的某些特点倾向于以微妙的方式影响受访者的答案（West and Blom 2016）。除了可能提高某类问题答案的准确度以外，采访者不参与数据采集还能大大降低成本（访谈时间是调查研究中最大的成本之一），并且增加了灵活性（受访者可以按自己的意愿随时参与调查，而不是受制于采访者的时间）。就挑战而言，如果调查是采访者管理的，那么采访者可以与受访者建立良好的关系，进而提高参与率，同时对受访者不理解的问题还能给予解释。对于问题特别多的调查问卷（可能会很乏味），采访者还可以保证受访者的完成度（Garbarski, Schaeffer, and Dykema 2016）。因此，从采访者管理的调查方式到计算机管理的调查方式，这种转变既带来了机遇也带来了挑战。

接下来，我将介绍两种提问的方法，表明研究人员如何借助数字时代的工具以不同的方式发问：用来在更合适的时间和地点测量内部状态的生态瞬时评估法（3.5.1小节）以及结合了开放式问题和封闭式问题优点的维基调查（3.5.2小节）。然而，由计算机管理的、不受地点限制的提问方式的出现，也意味着我们需要设计出受访者更喜欢的提问方式，这一过程有时被称为游戏化（3.5.3小节）。

3.5.1 生态瞬时评估法

研究人员可以分解大型的调查，然后将其融入人们的生活。

生态瞬时评估法将传统的调查分解，然后将其融入参与者的生活。因此，研究人员可以在合适的时间和地点进行提问，而不是在事情发生数周后才通过一个长时间的访谈来了解。

生态瞬时评估法主要有4个特征：（1）在现实环境中搜集数据；（2）评估的是个体当前或最近的状态或行为；（3）评估可能是基于事件的、基于时间的或随机引发的（取决于研究问题）；（4）随着时间的推移需进行多次评估（Stone and Shiffman 1994）。一天中人们可以不断通过智能手机进行交流，这大大提高了生态瞬时评估法的便利性。此外，智能手机上装有各种传感器，例如GPS（全球定位系统）和加速计，因此研究人员可以通过用户的活动情况启动相应的测量。例如，可将以智能手机设置为当受访者进入某特定街区时便向其提一个调查问题。

内奥米·杉江（Naomi Sugie）的研究可以很好地说明生态瞬时评估法的前景。自20世纪70年代以来，美国的监禁人数开始急剧上升。截至2005年，每10万美国人中就有约500人在狱中，这一比例要高于世界上其他任何地方（Wakefield and Uggen 2010）。入狱人数的激增也导致了出狱人数的激增，每年约有70万人出狱（Wakefield and Uggen 2010）。这些人出狱后面临着严峻的挑战，不幸的是，许多人最后又回到了监狱。为了了解和减少累犯，社会科学家和决策者需要了解这些人重新进入社会后的经历。然而，这些数据很难用标准的调查方法来搜集，因为这些曾经是罪犯的人往往是很难了解的，而且他们的生活非常不稳定。每隔几个月进行一次调查的测量方法会遗漏掉他们生活中大量的动态（Sugie 2016）。

为了更精确地研究他们重新进入社会的过程，杉江从新泽西州纽瓦克市所有出狱的人中抽取了一个131人的标准概率样本。她为样本中的每位参与者提供了一部智能手机，进而创建了一个既可以记录行为又可以提问的丰富的数据采集平台。杉江利用手机开展了两类调查。首先，她在上午9点和下午6点之间随机选了一个时间向参与者发送“体验抽样调查”，询问参与者当下的活动和感受。然后，在晚上7点，她会向参与者发送一个“每日调查”，询问他们当天的所有活动。除了这些调查问题以外，手机还会定期记录他们的地理位置，并以加密的方式记录有关打电话和发短信的元数据。通过将提问与观察相结合，杉江获得了这些人重新进入社会后详细的、高频的测量数据。

研究人员相信，找到稳定的、高质量的工作有助于人们成功地重返社会。然而，杉江发现，平均来说，其研究参与者找到的工作都是非正式的、临时的和零散的，但该平均描述掩盖了重要的异质性。杉江在其样本中发现了4个完全不同的群体：“早期退出”（最开始找过工作，但后来退出了劳动力市场）、“持续寻找”（融入社会前的大部分时间都花在找工作上）、“循环工作”（融入社会前的大部分时间都花在工作上）以及“低响应”（不会定期回答调查问题）。其中“早期退出”这一群体最开始找过工作，但后来没找到就放弃了，该群体尤其重要，因为他们可能是成功融入社会概率最低的群体。

人们可能会认为，出狱后找工作是一个很艰难的过程，这些人可能会因为沮丧而退出劳动力市场。因此，杉江通过她的调查还搜集了有关参与者情绪状态（一种通过行为数据难以评估出来的内部状态）的数据。令人惊讶的是，“早期退出”这一群体并没有称自己压力过大或过于悲伤，反倒是那些失败后继续找工作的人称自己过于忧虑悲伤。所有这些有关出狱人员行为和情绪状态的细微的、纵向的详细数据，对于理解他们所面临的阻碍以及降低他们重返社会的难度有着重要意义。但如果使用标准的调查，这些细微的数据就会被遗漏。

杉江的数据采集针对的是一个弱势群体，其数据采集方式可能会引发一些道德伦理方面的担忧。但杉江预先就考虑到了这些担忧，并在设计过程中采取了应对措施（Sugie 2014, 2016）。她所在大学的机构审查委员会作为第三方审查了她的数据采集程序，认为该程序符合所有现存规则。此外，杉江的方法与我在第6章所提倡的基于原则的方法相一致，在符合现有法规方面远远超出了要求的范围。例如，她获得了所有参与者的知情同意，这是很有意义的，她还同意参与者暂时关闭位置追踪，并且竭尽全力去保护她所搜集的数据。除了采用适当的加密技术和数据存储外，她还申请并获得了联邦政府的保密证书，这意味着她不会被迫将数据交给警察（Beskow, Dame, and Costello 2008）。因为考虑周全，所以我认为杉江的项目给其他研究人员提供了一个有价值的参考。尤其是她没有不加思考就行动而让自己陷入道德伦理的泥潭，也没有因为道德伦理上的复杂性而回避重要的研究。相反，她仔细思考，寻求合理建议，尊重她的参与者，并采取措施降低其研究的风险、增加研究的益处。

我认为从杉江的研究中可以学到以下三点：首先，提问的新方法与传统的抽样法是完全相容的，杉江就是从定义明确的抽样框总体中抽取了一个标准的概率样本。其次，高频、纵向的测量数据对于研究不规则的、动态的社会经历是很有价值的。最后，当调查数据采集与大数据资源相结合时（我认为这会越来越常见，我将在本章后面部分进行论述），就可能引发额外的道德伦理问题。我将在第6章更详细地探讨研究中的伦理问题，但杉江的事例表明，细心负责、考虑周密的研究人员是可以解决这些问题的。

3.5.2 维基调查

维基调查为封闭式问题和开放式问题的结合提供了新的可能。

除了能让我们在更合适的时间和更自然的环境中进行提问，新技术还让我们能够改变问题的形式。大多数调查问题都是封闭的，受访者只能从研究人员给定的几个选项中进行选择。一位著名的调查研究人员称该过程为“将单词放入人们的口中”。例如，以下就是一个封闭的调查问题：

下面一道题是有关工作的。请看这些条目，你能告诉我以下哪一项是你在一份工作中最看重的吗？

1. 高薪酬；

2. 没有被解雇的危险；

3. 工作时间短，有很多空闲时间；

4. 晋升机会；

5. 这份工作是很重要的，给人一种成就感。

但这些是全部可能的答案吗？研究人员将答案限制在这5个选项中会不会遗漏了一些重要的东西呢？与封闭式问题相对应的是开放式问题，以下是以开放的形式对同一个问题进行提问：

下面一道题是有关工作的。对于工作，人们追寻的是不同的东西。那你在一份工作中最看重的是什么呢？

尽管这两个问题看起来很相似，但霍华德·舒曼（Howard Schuman）和斯坦利·普雷瑟（Stanley Presser）的一项调查实验表明，它们可以产生非常不同的结果：近60%的以开放方式提问而搜集到的答案，都不在研究人员给定的选项中（图3.9）。

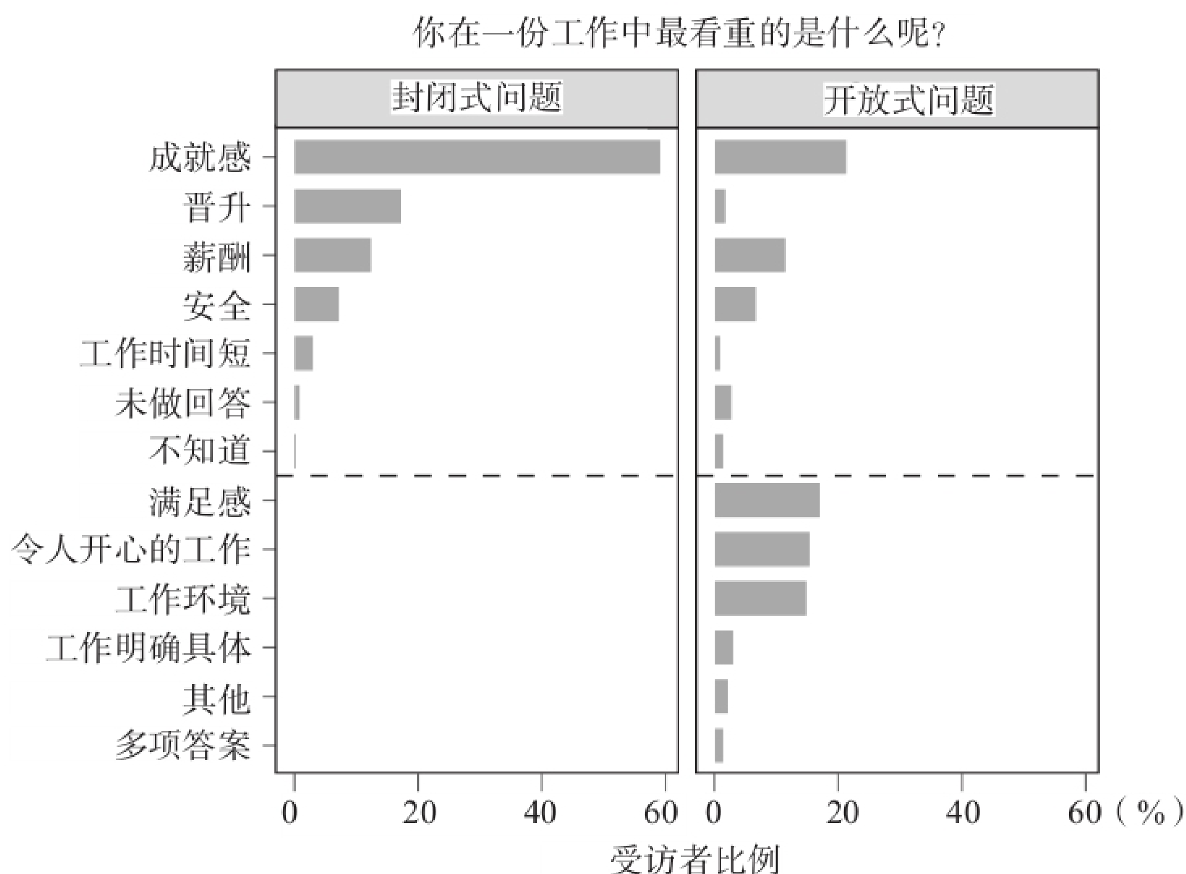


图3.9 一项调查实验的结果表明，采用封闭的方式提问与采用开放的方式提问所搜集到的答案不一样。改编自Schuman and Presser（1979），表1。

尽管开放式问题和封闭式问题可以产生完全不同的信息，而且两种形式的问题在调查研究的早期都很受欢迎，现在处于主导地位的却是封闭式问题。这并不是因为封闭式问题被证明能产生更好的测量数据，而是因为封闭式问题使用起来要简单很多，因为分析开放式问题的过程易于出错且成本高昂。研究人员逐渐不再采用开放式问题，这真是令人遗憾，因为正是那些研究人员事先不知道的信息才是最有价值的信息。

然而，从人类管理的调查到计算机管理的调查，这一转变为这个老问题找到了一个新的解决办法。如果我们现在能设计出融合了开放式问题和封闭式问题各自优点的调查问题，会怎么样呢？也就是说，

如果我们的调查既能搜集到新的信息又能保证答案易于分析，会怎么样呢？这正是卡伦·利维（Karen Levy）和我已着手想要实现的。

具体而言，卡伦和我认为，搜集和管理用户生成内容的网站可能会影响新型调查的设计。尤其是维基百科（内容主要由用户生成的动态开放系统的绝佳案例），它让我们很受启发，因此我们称这个新型调查为维基调查。正如维基百科会基于参与者的想法逐步发展，我们也设想了一个会基于参与者想法而逐步发展的调查。卡伦和我认为，维基调查应满足三个特性：贪婪性、协作性和适应性。然后，我们和一组网站开发人员一起创建了一个可以开展维基调查的网站：http://www.allourideas.org。

我们可以通过与纽约市长办公室共同开展的一个项目来了解维基调查的数据采集过程。该项目旨在将居民的想法整合到纽约市可持续发展规划中去。首先，市长办公室根据他们之前的外展服务（例如“要求所有大型建筑都要进行一定的能效升级”以及“把教孩子环保知识作为学校课程的一部分”）列出了25个想法，以此作为这样一个问题的备选答案：你认为哪一项更有利于创建一个更环保的、更好的纽约市？然后，计算机会随机从备选答案中抽取2个（例如“开放纽约市所有学校的操场作为公共体育场”和“增加哮喘发病率高的社区的植树量”），供受访者选择（图3.10）。受访者做出选择后，计算机会立即再随机抽取2个想法供其选择。

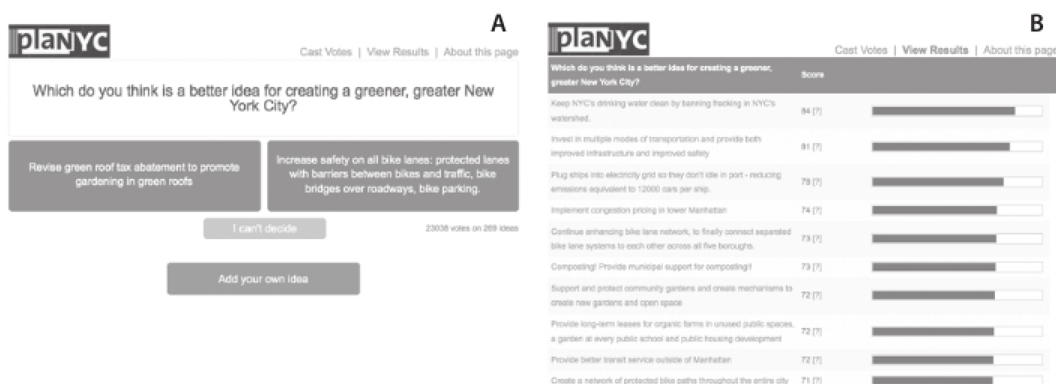


图3.10 一个维基调查的界面。图左是回答界面，图右是结果界面。经允许转载自Salganik and Levy（2015），图2。

只要受访者愿意，通过二选一或选择“我无法决定”，受访者可以一直回答他更偏向于哪种方案。最重要的是，受访者可以在任何时候贡献他们自己的想法，这些想法经过市长办公室的同意后，可以作为备选答案呈现给其他受访者。因此，受访者回答的问题既是开放的又是封闭的。

为了搜集居民的反馈信息，市长办公室于2010年10月启动了该维基调查，同时还开展了一系列的社区会谈。在大约4个月的时间里，1436名受访者贡献了31893个答案以及464个新想法。至关重要的是，前10个最受欢迎的想法中有8个是受访者提出来的，而不是源于市长办公室起初列出的25个想法。并且，正如我们在论文中所描述的那样，受访者贡献的想法比研究人员给定的想法更受欢迎的现象在许多维基调查中都出现过。换句话说，通过允许受访者贡献自己的想法，研究人员能够了解到更多的信息，而这些信息在使用更封闭的方法进行调查时可能会被遗漏掉。

除了这些具体调查的结果以外，我们的维基调查项目还展示了数字研究的成本结构是如何让研究人员能以不同的方式接触世界的。现在，学术研究人员能够创建可供许多人使用的真实系统：我们已经主办了超过10000项维基调查，搜集了1500多万份答案。这种创造可以大规模使用的东西的能力源于这样一个事实：一旦一个网站建立起来了，那么让世界上的每个人都免费使用它基本上不会再产生成本（当然，如果我们采取由人类管理的访谈，就无法实现这一点）。此外，这样的规模可以使研究人员开展不同类型的研究。例如，这1500多万份答案以及大量的参与者为未来的方法研究提供了一个宝贵的测试场。在第4章介绍实验时我将进一步描述数字时代成本结构（尤其是成本不会随着所搜集数据量的增加而增加）所创造的其他研究机会。

3.5.3 游戏化

标准调查对参与者来说是很无聊的。这种情况可以改变，而且必须改变。

到目前为止，我已经向大家介绍了一些新的提问方法，而计算机管理的调查则对这些方法的出现起到了促进作用。但计算机管理的调查也有一个缺点，那就是没有一个采访者来帮助诱导和保持受访者的参与度。这之所以成为一个问题，是因为许多调查既费时又枯燥。因此，在未来，调查设计者在设计过程中将不得不考虑参与者的想法，以便使回答问题的过程更愉快、更像游戏。这一过程有时被称为游戏化。

我将通过“朋友感觉”（Friend Sense）这项调查来说明怎样才可能做出一项有趣的调查。该调查是在脸谱网上进行的，看起来像是一个游戏。沙拉德·戈埃尔、温特·梅森（Winter Mason）和邓肯·瓦茨旨在通过这项调查来评估人们认为自己与朋友有多相似，以及实际上与朋友有多相似。这个有关真实的态度相似度和感知的态度相似度的问题，可以直接反映人们精确感知自己社交环境的能力，并对政治极化和社会变化的动态产生影响。从概念上讲，真实的态度相似度和感知的态度相似度是很容易测量的。研究人员可以就某方面的观点向人们提问，然后再问他们的朋友（这样可以测得真实的态度相似度），也可以让人们猜其朋友的态度（这样可以测得感知的态度相似度）。但可惜的是，既采访受访者又采访其朋友，实施起来特别困难。因此，戈埃尔和同事把他们的调查变成了一个好玩的脸谱网应用程序。

在一位参与者同意参加一项研究后，该应用程序会从该参与者的脸谱网中选择一位好友，然后就该好友的态度向参与者提问（图3.11）。在回答有关随机选择的朋友的问题时，该参与者也要回答有

关自己的问题。在答完有关一个朋友的问题时，系统会告诉该参与者其答案是否正确，如果该参与者的朋友没有作答，该参与者还可以鼓励他作答。因此，这项调查在一定程度上是通过病毒式招募来传播的。

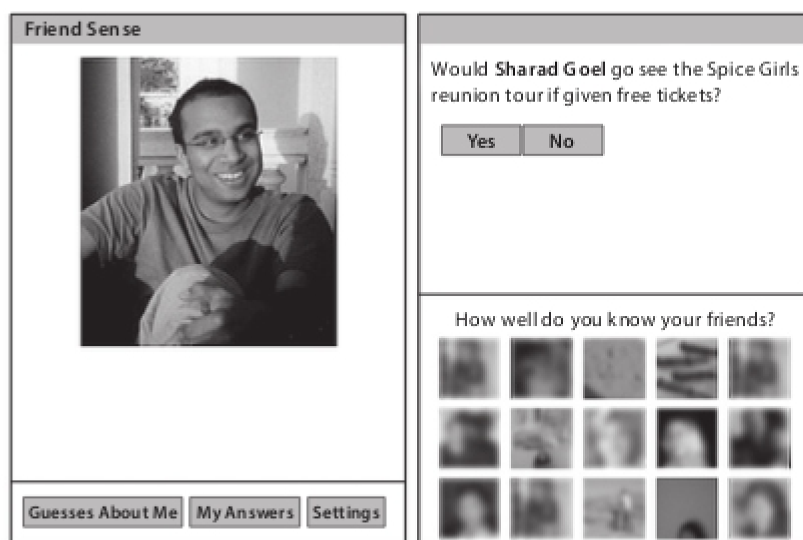


图3.11 “朋友感觉”的界面（Goel, Mason, and Watts 2010）。研究人员把标准的态度调查变成了一项有趣的、类似游戏的体验。应用程序向参与者提的问题有严肃的也有轻松的。好友头像经过了模糊处理。经沙拉德·戈埃尔允许转载。

这些有关态度的问题改编自美国综合社会调查。例如：“在中东局势中，相比于巴勒斯坦人，（你的朋友）更同情以色列人是吗？”以及“（你的朋友）会为了让政府能够提供全民医保而缴更多的税吗？”除了这些严肃的问题以外，研究人员还会设置更轻松的问题：“相比于啤酒，（你的朋友）更喜欢葡萄酒是吗？”以及“（你的朋友）更希望拥有读心术而不是会飞是吗？”这些更轻松的问题会让参与者觉得这个过程很有趣，同时也让我们能够进行一项有趣的比较：参与者和朋友在严肃的政治问题上以及有关喝酒和超能力的轻松问题上的态度相似度会基本一样吗？

这项研究主要得出了三个结论。首先，相比于陌生人，朋友更可能给出相同的答案，但即使是很亲密的朋友，也在约30%的问题上持不同的观点；其次，参与者往往高估自己与朋友的相似度，换句话说，

朋友之间在看法上的大多数差异都没有被注意到；最后，在有关政治的严肃问题上以及有关喝酒和超能力的轻松问题上，参与者对自己与朋友在看法上的差异的感知基本是一样的。

尽管这款应用程序现在已经不能再玩了，但它很好地说明了研究人员如何能让一个标准的态度调查变得有趣。更广泛地说，通过一些创造性的想法和设计工作，研究人员就有可能改善调查参与者的用户体验。因此，下次你设计一项调查时，要花点时间思考一下你能做些什么来让你的参与者感觉更好。有些人可能会担心这些追求游戏化的举措会影响数据质量，但我认为，觉得调查无聊的参与者对数据质量的影响要大得多。

戈埃尔和同事的调查研究也体现了下一节的主题：将调查与大数据资源结合起来。在戈埃尔和同事的事例中，他们通过将调查与脸谱网结合起来，自动获得了参与者的好友列表。在下一节中，我们将更详细地探究调查与大数据资源之间的结合。

3.6 与大数据资源相结合的调查

将调查与大数据资源结合起来，能让你得出单独通过调查或大数据资源所无法得出的评估结论。

大多数调查都是独立进行的。它们没有以彼此为基础，也没有借助世界上现有的其他数据。这种现象将会改变。将调查数据与第2章介绍的大数据资源结合起来，我们将得到更多益处。通过这两种类型数据的结合，我们就能做那些只通过调查数据或只通过大数据资源不可能做到的事情。

将调查数据与大数据资源结合起来有几种不同的方法。在本节中，我将介绍两种有用而截然不同的方法，我称它们为丰富型提问和扩充型提问（图3.12）。虽然对每种方法我都将通过一个详细的事例来说明，但大家应该可以看出，这两种方法其实可以被用于不同类型的调查数据和不同类型的大数据。此外，大家还应注意，这两个事例中的每一个都可以用两种不同的视角看待。回想一下第1章的内容，有些人会把这些研究看作“非现成”调查数据增强“现成”大数据的事例，而其他人则会把它们看作“现成”大数据增强“非现成”调查数据的事例。大家应该兼备这两种视角。最后，大家要注意这两个事例如何说明了调查数据和大数据资源应该彼此互补而不是替代。

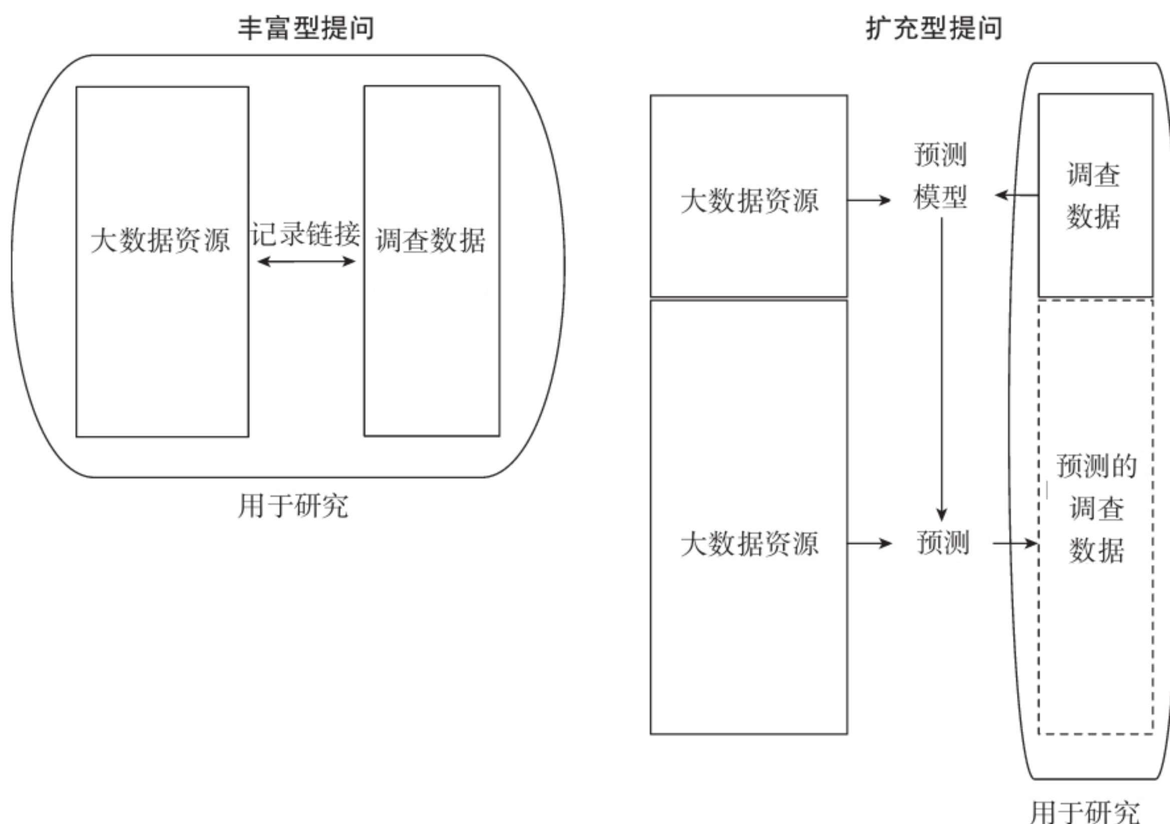


图3.12 两种将大数据资源和调查数据结合起来的主要方法。在丰富型提问（3.6.1小节）中，大数据资源中含有研究人员感兴趣的核心数据，而调查数据围绕该大数据资源构建起了必要的背景。在扩充型提问（3.6.2小节）中，大数据资源中没有研究人员感兴趣的核心数据，但研究人员可以用它来扩充调查数据。

3.6.1 丰富型提问

在丰富型提问中，大数据资源含有一些重要的测量数据，但缺失其他一些测量数据，而调查数据围绕该大数据资源构建起了必要的背景。

将调查数据和大数据资源结合起来的一种方法，我称之为丰富型提问。在丰富型提问中，大数据资源中含有一些重要的测量数据，但缺失其他一些测量数据，因此研究人员需要通过一项调查来搜集这些缺失的数据，然后将两部分数据资源结合起来。丰富型提问的一个事

例是我在3.2节中提到的伯克和克劳特针对脸谱网上的互动是否会增进友谊所开展的研究。在该研究中，伯克和克劳特把调查数据与脸谱网的日志数据进行了结合。

然而，伯克和克劳特当时的工作环境意味着他们无须面对那些进行丰富型提问的研究人员通常会面临的两大难题。其中一个是，如果两个数据资源中都没有可用来确保一个数据集中的正确记录与另一个数据集中的正确记录相匹配的唯一标识符，那么就很难将个体层面的数据集链接起来（这是一个被称为记录链接的过程）。第二个难题是，大数据资源的质量通常很难评估，因为数据创建的过程可能是不对外公开的，并且大数据资源容易受到第2章所描述的问题的影响。换句话说，丰富型提问经常需要将调查与质量未知的黑匣子似的数据资源链接起来，而这一过程很容易出错。尽管存在这些问题，但我们还是可以利用丰富型提问开展重要的研究，正如斯蒂芬·安索拉比赫（Stephen Ansolabehere）和埃坦·赫什（Eitan Hersh）针对美国的投票模式所进行的研究那样。

投票率一直是大量政治科学研究的主题，而且在过去，研究人员对于谁投票以及为什么投票的理解基本上都基于对调查数据的分析。但在美国，投票是一种非比寻常的行为，因为政府会记录每个公民是否投票。（当然，政府没有记录每个公民把票投给了谁。）多年来，政府的这些投票记录都是纸质版的，分散在全国各地的地方政府办公室中。这使政治科学家很难（但也不是不可能）获得全体选民的投票记录，并将他们在调查中关于投票所说的内容和实际的投票行为进行比较（Ansolabehere and Hersh 2012）。

但这些投票记录现在已经被数字化了。通过系统地搜集和汇总这些记录，一些私人公司已经创建了包含所有美国人投票行为的全面主投票文件。安索拉比赫和赫什就选择了与其中一家公司Catalist（凯利板）合作，以利用其主投票文件帮助他们更好地了解全体选民。此

外，因为他们的研究依赖于上述这家公司（该公司在数据采集和汇总方面投入了大量的资源）所搜集和管理的数字记录，所以他们现在要比之前没有公司帮助且使用模拟记录开展研究时多了许多优势。

像第2章的许多大数据资源一样，安索拉比赫和赫什获得的主投票文件中也没有太多他们所需要的人口统计、态度以及行为方面的信息。事实上，他们特别感兴趣的是比较调查中报告的投票行为和经过验证的投票行为（即Catalist数据库中的信息）。因此，安索拉比赫和赫什借助本章前面提到的大型社会调查——合作国会选举研究，搜集了他们想要的数据。然后，他们把搜集来的数据交给了Catalist，Catalist汇总后又将包括经验证的投票行为（源于Catalist）、自我报告的投票行为（源于合作国会选举研究）以及受访者的人口统计资料和态度在内的数据文件返回给了他们（图3.13）。换句话说，安索拉比赫和赫什的研究只有将投票记录与调查数据结合起来才能开展，如果只有投票记录或只有调查数据，研究是不可能开展的。

利用结合后的数据文件，安索拉比赫和赫什得出了三个重要结论。首先，过度报告投票行为的现象很是普遍：未投票者中几乎有一半的人报告称自己投过票，并且如果有人报告称自己投过票，实际上他真正投过票的概率只有80%。其次，过度报告并不是随机的：过度报告在高收入、受过良好教育、参与公共事务的党派人士中更为常见。换句话说，最有可能投票的人也最有可能谎报自己投过票。最后，也是最重要的一个结论是，由于过度报告的系统性，投票者和未投票者之间的实际差异比调查所显示的要小。例如，拥有学士学位的人报告称自己投过票的可能性要比没有的人高约22%，而其实际投票的可能性只高出了10%。事实证明，相比于预测谁会真正投票，现有的以数据源为基础的理论在预测谁会报告称自己投过票（这也是研究人员过去所使用的数据）方面，准确度会更高。因此，安索拉比赫和赫什的实证发现表明，我们需要新的理论来理解和预测投票。

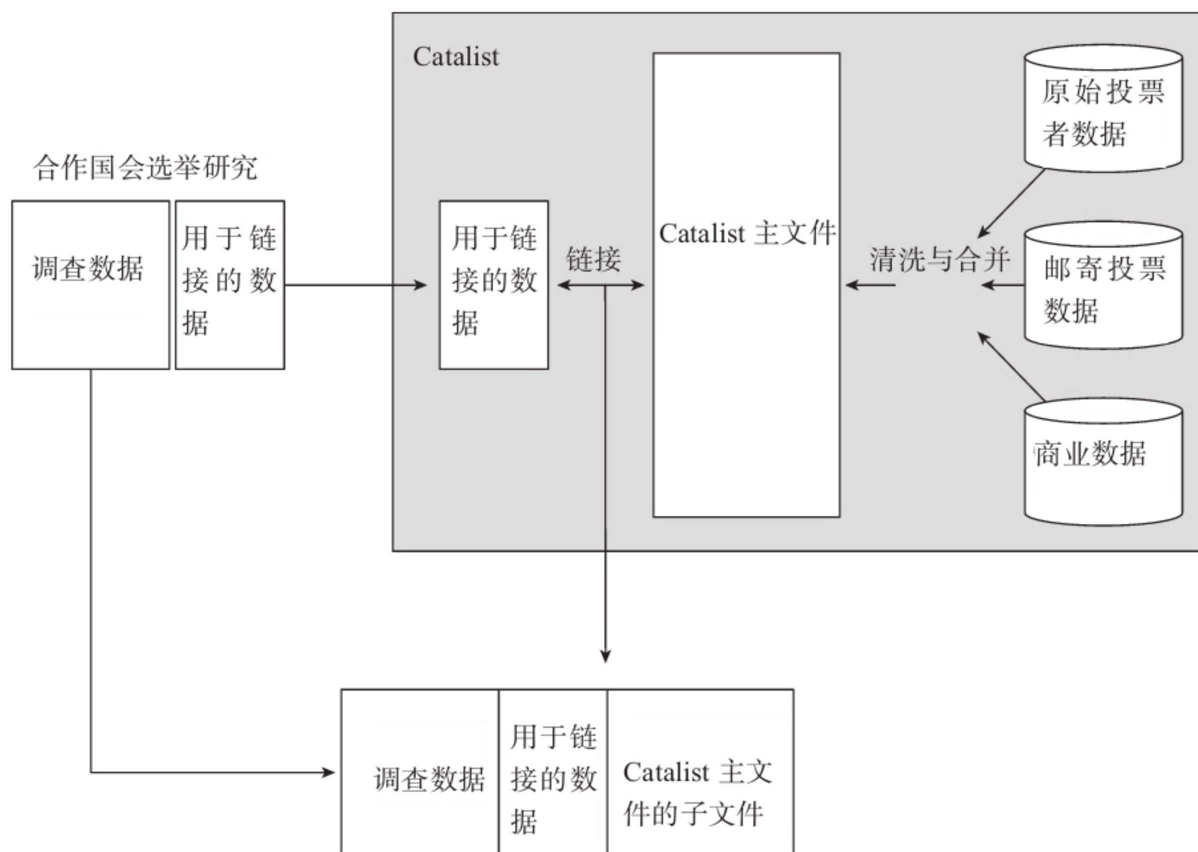


图3.13 安索拉比赫和赫什的研究示意图。为了创建主数据文件，Catalist需将多个不同来源的信息进行汇总和统一。这一过程，无论多么小心，都会使原始数据资源中的错误得以传播，同时还会引发新的错误。第二个错误的来源是调查数据和主数据文件之间的记录链接。如果每个人在上述两个数据资源中都有一个稳定的、唯一的标识符，那么链接就会很简单。但Catalist只能通过不完美的标识符（在该事例中是指姓名、性别、出生年份以及家庭住址）进行链接。不幸的是，在许多情况下会出现不完整或不精确的信息，例如一位名叫荷马·辛普森（Homer Simpson）的投票者可能会被登记为荷马·杰·辛普森、荷马·J. 辛普森，甚至是拼错单词。尽管主数据文件中和记录链接过程中都有可能出现错误，但通过几种不同类型的检查，安索拉比赫和赫什对他们的评估结论还是有信心的。

但对于这些结果我们又能信任多少呢？不要忘了这些结果依赖于一个易于出错的链接过程，而且这一过程需要链接至拥有未知数量错误的黑匣子似的数据中。更具体地说，这些结果取决于两个关键步骤：（1）Catalist将许多完全不同的数据资源汇总后，形成一个精确的主数据文件；（2）将调查数据链接至上述主数据文件。其中每个步

骤都很困难，而且任何一个步骤中的错误都会导致研究人员得出错误的结论。

然而，作为一个公司，数据处理和链接对其继续生存来说至关重要，所以Catalist能以学术研究人员无法企及的规模投入资源，解决这些问题。在他们的论文中，安索拉比赫和赫什通过一系列步骤检查了上述两个步骤的结果（尽管有些是不对外开放的），这些检查对其他想要将调查数据和黑匣子似的大数据资源结合起来的研究人员可能会有所帮助。

一般来说，研究人员通过该研究能学到什么经验呢？首先，无论是利用调查数据丰富大数据资源还是利用大数据资源丰富调查数据（我们可以用任何一种方式看待该研究），都有巨大的价值。通过将两个数据资源相结合，研究人员就能做那些如果只有调查数据或只有大数据资源就不可能做的事情。其次，尽管汇总后的商业数据，例如Catalist的数据，不应被看作绝对真实，但在有些情况下，这些数据也是有用的。怀疑论者有时会将这些汇总后的商业数据与绝对真实进行比较，并指出这些数据资源存在着不足。但就这一情况而言，怀疑论者其实正在做错误的比较，因为研究人员使用的所有数据都达不到绝对真实。因此，比较好的做法是，将汇总后的商业数据资源与其他同样会有错误的可用数据资源（例如，受访者自我报告的投票行为）进行比较。最后，在某些情况下，研究人员可能会从许多私人公司在搜集和统一复杂的社会数据集方面的巨大投资中获益。

3.6.2 扩充型提问

扩充型提问会通过一个预测模型将源于少数人的调查数据与源于许多人的一个大数据资源结合起来。

将调查数据和大数据资源结合起来的另一种方法，我称之为扩充型提问。在扩充型提问中，研究人员会通过一个预测模型将少量的调查数据与一个大数据资源结合起来，然后利用结合后的数据得出评估结论，这些评估结论的规模或粒度是只通过调查数据或大数据资源不可能实现的。扩充型提问的一个重要事例是乔舒亚·布卢门斯托克的研究，他想搜集有助于指导贫穷国家发展的数据。在过去，搜集此类数据的研究人员一般只能采取以下两种方法中的一种：抽样调查或人口普查。只需要研究人员采访少量人的抽样调查比较灵活、及时且成本相对较低，但因这些调查是基于一个样本的，所以其分辨率通常是有限的。也就是说，通过一项抽样调查，通常很难对特定的地理区域或人口群体做出评估。而人口普查则试图采访每一个人，因此研究人员可以通过人口普查对小的地理区域或人口群体进行评估。但人口普查往往成本高、关注面小（它只包含少量的问题），而且不及时（它按固定的时间表进行，例如每隔10年普查一次）（Kish 1979）。与其勉强接受抽样调查或是人口普查，研究人员还不如设想一下能否将两种方法各自的最佳特征结合起来，能否每天向每一个人提出每一个问题。显然，这种无处不在的、不间断的调查只是一种社会科学幻想。但通过将源于少量人的调查问题与源于许多人的数字痕迹相结合，我们似乎确实可以接近这一幻想。

布卢门斯托克的研究始于他与卢旺达最大的手机供应商的合作，该供应商向其提供了2005—2009年间约150万名用户的匿名记录。这些记录含有每次通话和每条短信的相关信息，例如起始时间、持续时间以及呼叫方和接收方大致的地理位置。在我谈论统计问题之前，值得指出的是，获取数据这第一步对许多研究人员来说可能是最难的步骤之一。正如我在第2章所描述的，大多数大数据资源都是研究人员难以获取的。电话元数据尤其难以获取，因为它们基本上是不可能被“匿名化”的，而且几乎一定会包含参与者认为敏感的信息（Mayer, Mutchler, and Mitchell 2016; Landau 2016）。在布卢门斯托克的研究中，研究人员在保护数据方面很是谨慎，而且有一个第三方（即

他们的机构审查委员会）负责监督他们的工作。我将在第6章更详细地探究这些道德伦理问题。

布卢门斯托克感兴趣的是衡量财富状况和幸福感，但通话记录中并没有直接记录这些特性。换句话说，对该研究来说，这些通话记录具有不完整性，这是大数据资源所共有的一个特征，在第2章中我详细介绍过这一点。但通话记录很可能包含一些能间接提供有关财富状况和幸福感的信息内容。鉴于这种可能性，布卢门斯托克想，是否有可能训练一个机器学习模型，该模型能通过通话记录预测某个人在一项调查中会如何作答？如果这是可行的，那么布卢门斯托克就能利用这一模型预测150万名用户的调查答案。

为了创建和训练这样一个模型，布卢门斯托克和来自基加利科学技术研究所（Kigali Institute of Science and Technology）的研究助理随机抽取了约1000名用户。研究人员向参与者解释说明了研究的目的，并就将调查答案与通话记录结合起来征求了他们的同意，然后向他们提了一系列的问题，以衡量其财富状况和幸福感，例如“你有收音机吗？”以及“你有自行车吗？”（部分问题列表参见图3.14）。所有参与调查的人都获得了经济报酬。

接下来，布卢门斯托克采用了机器学习中常见的两步法：特征工程、监督式学习。首先，在特征工程这一步中，布卢门斯托克将通话记录转换成了关于每个受访者的一组特征，数据科学家可能会称这些特征为“特征”，而社会科学家则可能称之为“变量”。例如，布卢门斯托克会计算出每个人的总活跃天数、联系过的人（不同的人）的数量以及通话费等。这一步中至关重要的一点是，好的特征工程需要研究人员了解研究环境。例如，如果区分国内电话和国际电话很重要（我们可能会认为打国际电话的人更富有），那么这一工作就必须在特征工程这一步进行。而对卢旺达知之甚少的研究人员可能就不会考虑这一特征，模型的预测性能就会受到影响。

图3.14 用通话记录训练统计模型的预测精度。改编自Blumenstock（2014），表2。

接下来，在监督式学习这一步中，布卢门斯托克创建了一个模型，根据每个人的特征预测其调查答案。在该步骤中，布卢门斯托克采用了逻辑回归，但其实他也可以采用其他统计或机器学习方法。

那么该模型的效果如何呢？根据从通话记录中提取的特征，布卢门斯托克是否能利用该模型预测用户对诸如“你有收音机吗？”以及“你有自行车吗？”的调查问题的回答吗？为了评估其预测模型的性能，布卢门斯托克采用了交叉验证，该方法在数据科学中很常用，但在社会科学中很少被采用。交叉验证的目的是公平评估一个模型的预测性能，做法是利用不同的子数据集训练并测试该模型。具体来说，布卢门斯托克先将其数据分成了10组，每组包含100人的数据。然后，他用其中9组数据来训练模型，然后用剩余的1组数据来评估该模型的预测性能。他将这一过程重复了10次，每次选1组不同的数据来验证模型的预测性能，而其余9组数据则用于训练模型，最后取平均值。

该模型对某些特征的预测精度是很高的（图3.14），例如预测某人是否有收音机的精确度能达到97.6%。这听起来可能很不错，但将一个复杂的预测模型与一个简单的替代方法进行比较通常是很重要的。在该事例中，一个简单的替代方法就是预测每个人都会给出最常见的回答。例如，97.3%的受访者回答说自己有收音机，因此，如果布卢门斯托克预测每个人都会回答说自己有收音机，那么他的精确度就是97.3%，这与他更复杂的预测模型的表现（97.6%的精确度）竟惊人地相似。换句话说，所有复杂的数据和建模工作只是把预测精确度从97.3%提高到了97.6%。但对其他问题，例如“你有自行车吗？”，预测精确度就从54.4%提高到了67.6%。更概括地说，图3.15表明，对某些特征来说，相比于简单的基线预测（即预测每个人都会给出最常见的回答），布卢门斯托克的模型并没有明显提高精确度，但对其他一些特征来说还是有些改善的。但仅从这些结果来看，你可能会觉得这种方法并不是特别有前景。

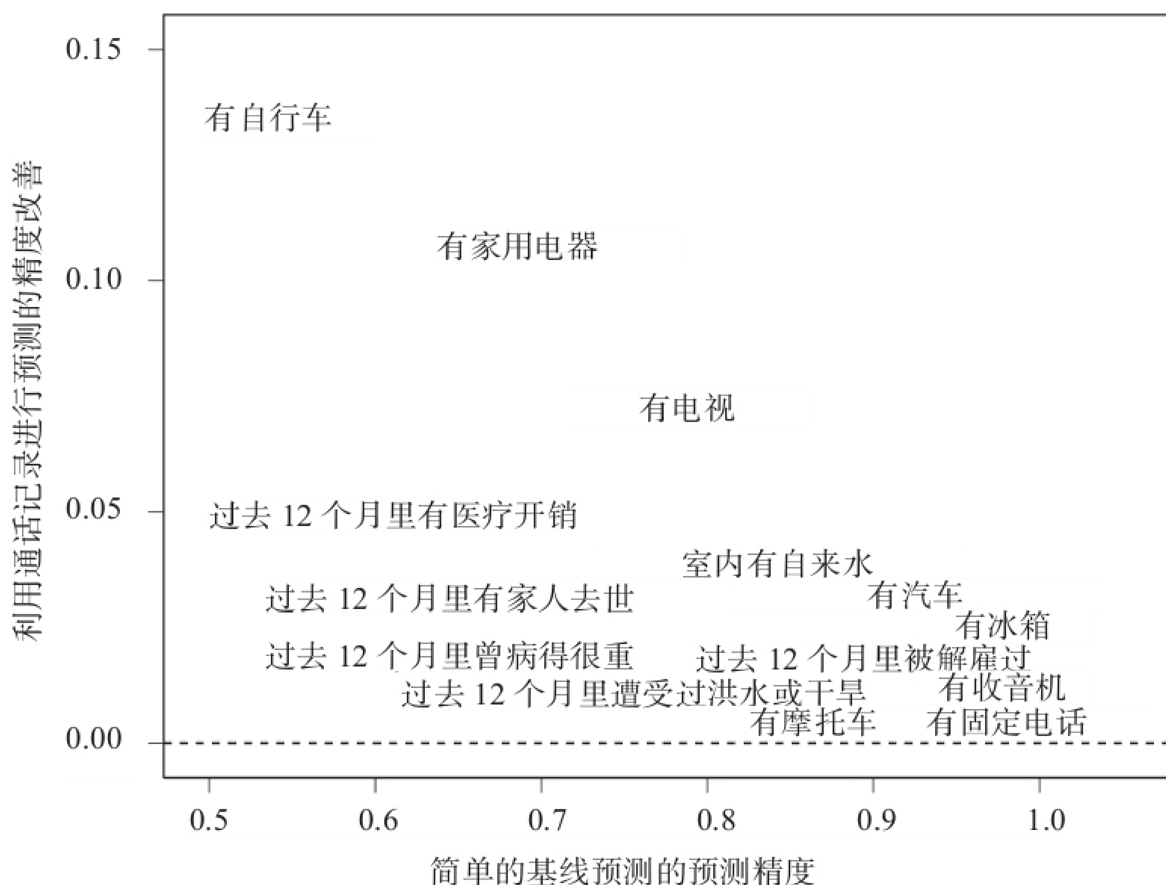


图3.15 利用通话记录训练的统计模型的预测精度与简单的基线预测的预测精度的对比。为了避免重叠，有的数值有轻微调整。改编自Blumenstock（2014），表2。

然而，仅仅一年后，布卢门斯托克和两位同事，加布里埃尔·卡达穆罗（Gabriel Cadamuro）和罗伯特·翁（Robert On），就大大改进了研究结果（Blumenstock, Cadamuro, and On 2015），并在《科学》杂志上发表了他们的论文。这一改进主要有两个技术原因：（1）他们采用了更复杂的方法（即在特征工程这一步中采用了新的方法，并创建了更复杂的模型来利用特征预测回答）；（2）他们不再试图推断单个调查问题（例如，“你有收音机吗？”）的答案，而是试图推断综合财富指数。这些技术上的改进意味着他们可以合理地利用通话记录预测样本中用户的财富状况。

但预测样本中用户的财富状况并不是他们研究的最终目标，他们的最终目标是将抽样调查和人口普查各自的最佳特征结合起来，从而

对发展中国家的贫穷状况进行准确的、高分辨率的评估。为了判断他们是否有能力实现这一目标，布卢门斯托克和同事用他们的模型和数据预测了150万名用户的财富状况。他们还利用通话记录中的位置信息（通话记录中有每次通话时用户离得最近的手机信号塔的位置）评估了每个人大致的居住地（图3.16）。通过将这两项评估结合在一起，布卢门斯托克和同事得出的评估结论，是关于用户财富地理分布的极细粒度（指空间粒度）的。例如，他们能够估算出卢旺达2148个街区中每一个街区的平均财富状况。

那么这些评估结论与该地区的真实贫困水平的符合程度如何呢？在回答这个问题之前，我想强调一个事实，那就是大家有很多理由持怀疑态度。例如，在个体层面上进行的预测，其结果相当参差不齐（图3.17）。此外，也许更重要的一点是，有手机的人与没有手机的人可能会有系统性的差别。因此，布卢门斯托克和同事可能也会受到覆盖面误差的困扰，类似于我之前所描述的1936年《文学文摘》的调查。

为了了解他们评估结论的质量，布卢门斯托克和同事需要将其数据与其他东西进行比较。幸运的是，就在他们进行该项研究的同时，另一组研究人员也正在卢旺达进行一项传统的社会调查。这项调查是广受重视的人口统计和健康调查的一部分，拥有大量预算，采用的是高质量的传统方法。因此，人口统计和健康调查的评估结论可以被合理地认为是黄金标准。人们将这两种评估进行比较后发现，它们非常相似（图3.17）。换句话说，通过将少量调查数据与通话记录结合起来，布卢门斯托克和同事得出了与采用黄金标准的方法所得出的评估结论相类似的结论。

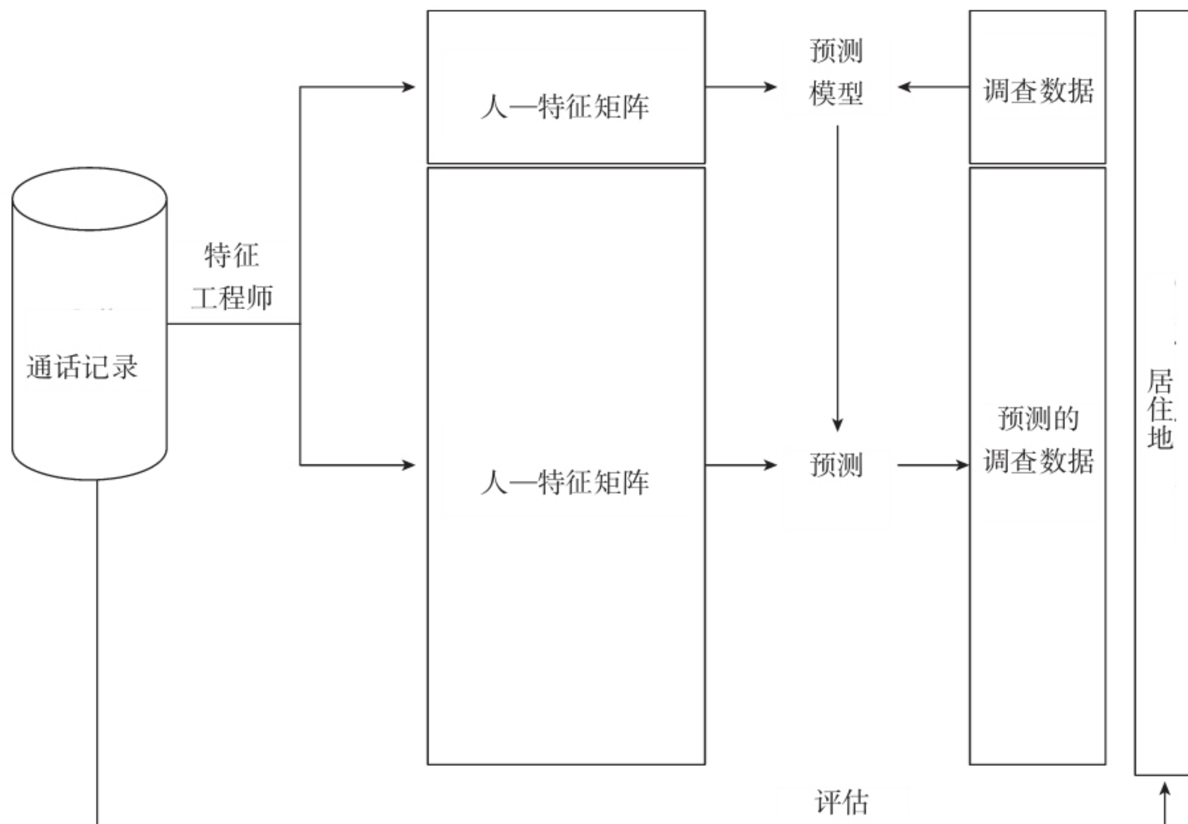


图3.16 布卢门斯托克、卡达穆罗以及翁的研究示意图。首先，他们将手机供应商的通话记录转换成了矩阵，其中每个人占一行，每个特征（即变量）占一列。接下来，他们创建了一个监督式学习模型，以通过上述矩阵预测调查答案。然后，他们利用该模型预测150万名用户的调查答案。此外，他们还根据这些用户打电话时的位置评估了150万名用户大致的居住地。在把这两项评估，即财富状况和居住地，结合起来后，其结果与人口统计和健康调查的评估结果很相似，而人口统计和健康调查被认为是黄金标准的传统调查。

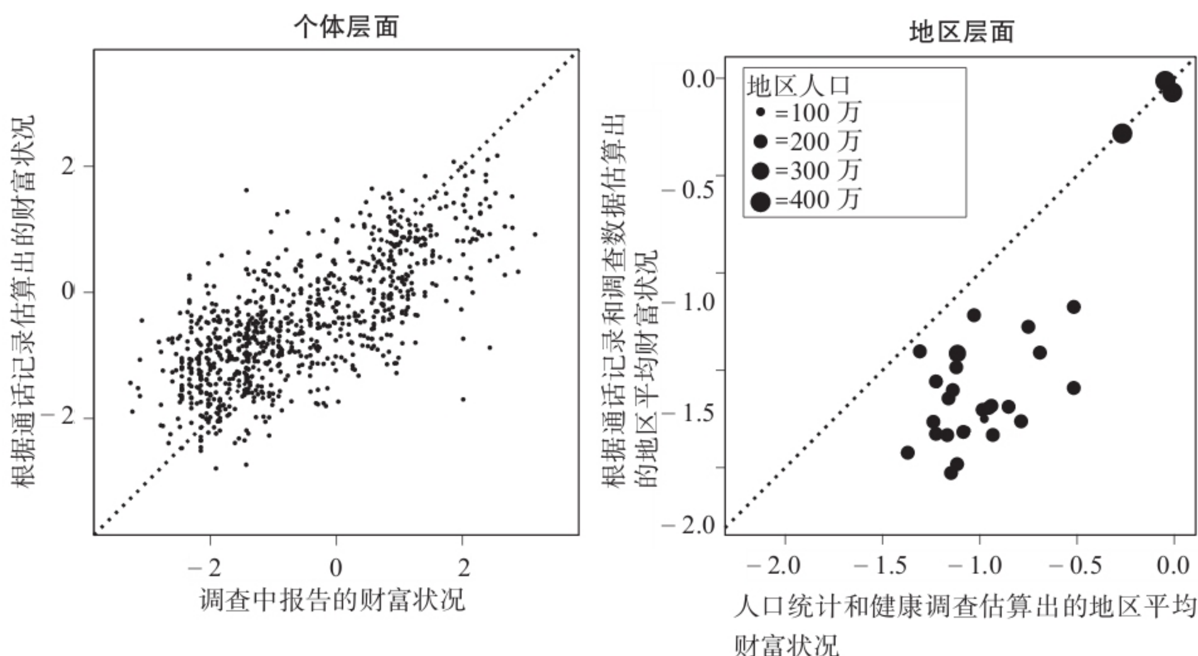


图3.17 布卢门斯托克、卡达穆罗以及翁的研究结果。在个体层面上，研究人员能通过某人的通话记录来合理预测其财富状况。基于个体层面的财富状况和居住地的评估所得出的卢旺达30个地区的地区级财富状况评估，与人口统计和健康调查的评估结果很相似，而人口统计和健康调查被认为是黄金标准的传统调查。改编自Blumenstock，Cadamuro，and On（2015），图1a和图3c。

持怀疑态度的人可能会认为这些结果是令人失望的。毕竟，对布卢门斯托克和同事的研究的一种解读方式是，利用现有的方法能更可靠地得出他们通过大数据和机器学习所得出的评估结论。但我认为这并不是解读该研究的正确方式，原因有以下两点。首先，相比于现有的方法，布卢门斯托克和同事得出评估结论所采用的方法要快10倍，成本降为1/50（以可变成本计算）。正如我在上文所论述的一样，研究人员忽视成本问题可能会给自己带来麻烦。就拿布卢门斯托克和同事的研究为例，相比于每隔几年开展一次（这是该调查的一个标准）的人口统计和健康调查，布卢门斯托克和同事在成本上的巨大优势意味着他们每个月都可以开展这样的研究，这将为研究人员和决策者带来诸多好处。其次，该研究采取的方法经调整后可以被用于许多不同的研究。该方法仅需要两类资源及两个步骤。这两类资源分别是：（1）广而薄的大数据资源（即该资源包含许多人，但没有你需要的关

于每个人的信息）；（2）窄而厚的调查数据（即该数据只包含少量人，但其中有你需要的关于这些人的信息）。然后需要用两步把这些资源结合起来。首先，为两个数据资源中的人创建一个机器学习模型，该模型可以利用数字痕迹预测调查答案。然后，用该模型推断上述大数据资源中每个人的调查答案。因此，如果你有一些问题想问很多人，那就可以找一个有关这些人的大数据资源，哪怕是你不感兴趣的大数据资源也行，因为它也许能预测这些人将如何回答你的问题。也就是说，布卢门斯托克和同事最初并不是关心通话记录，他们之所以留意这些通话记录，是因为其可以预测他们真正关心的调查答案。扩充型提问有别于我之前描述过的嵌入式提问，你只需要对大数据资源有非直接的兴趣即可。

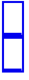
综上所述，布卢门斯托克通过采用扩充型提问得出的结论与符合黄金标准的调查评估结论类似。该研究事例也阐明了扩充型提问和传统的调查方法各自存在的利弊。利用扩充型提问进行评估更及时，具有明显的成本优势，且得出的结论粒度更细。但这类扩充型提问目前还没有很强的理论基础。仅凭这一个事例并不能说明该方法何时有效、何时无效，而且使用该方法的研究人员需格外注意因所使用的大数据资源包含某些人而未包含另外一些人而可能导致的偏差。此外，扩充型提问目前还没有好的方法来量化其评估结论的不确定性。幸运的是，扩充型提问与统计学中的三大领域有着很深的关联。这三大领域分别是小区域估计（Rao and Molina 2015）、填补法（Rubin 2004）以及基于模型的事后分层（该领域与我在前面介绍的“P先生”这一方法紧密相关）。鉴于这些很深的关联性，我预计扩充型提问的许多方法论基础都将很快得到增强。

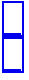
最后，通过对比布卢门斯托克的第一次和第二次尝试，我们还能学到有关数字时代社会研究的一个重要经验：开始并不是结束。也就是说，许多时候，第一次的方法可能并不是最好的，但如果研究人员继续努力，情况就会变得更好。更广泛地说，在评估数字时代社会研

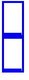
究的新方法时，进行以下两项截然不同的评估是很重要的：（1）该方法在当下的效果如何？（2）随着数据概况的改变以及研究人员投入更多的关注在这个问题上，该方法的效果又将如何？尽管研究人员接受过第一类评估的培训（评估一项特定研究的好坏），但第二类评估往往更为重要。

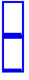
3.7 结论

从模拟时代到数字时代的转变正在为调查研究人员创造新的机会。在本章中，我提出大数据资源不会取代调查，而且其丰富性还将提升而不是降低调查的价值（3.2节）。然后，我总结了在调查研究的前两个时代发展起来的调查误差总框架，该框架有助于研究人员开发和评估调查研究第三个时代的方法（3.3节）。我预计会出现令人兴奋的机会的三个领域分别是：（1）非概率抽样（3.4节），（2）计算机管理的调查（3.5节），（3）将调查和大数据资源结合起来（3.6节）。技术和社会方面的变化驱动着调查研究不断地向前发展。我们应该拥抱这一发展趋势，并继续从之前的时代汲取智慧。









4.4 超越简单实验

让我们从以下三个概念入手来实现从简单实验到丰富实验的突破：效度、处理效应的异质性和原理。

刚开始做实验的研究者往往会把注意力集中在一个具体的、范围狭小的问题上：这个处理能“起作用”吗？例如，一个志愿者打来的电话能促使一个人投票吗？将网站按钮从蓝色换成绿色能增加广告的点击率吗？不过遗憾的是，对于“起作用”的不严谨的措辞掩盖了这样一个事实，即研究范围狭隘的实验是无法真正反映一个处理是否能够在一般意义上“起作用”的。相反，该类型实验其实是为了回答一个更具体的问题：在特定的时间对特定数量的参与者进行此次特定的实验，会产生怎样的平均效应呢？我个人会把以这种范围狭小的问题作为出发点的实验称为简单实验。

简单实验能够提供有价值的信息，但无法回答许多重要且有趣的问题，例如是否有人会对某一处理表现出比别人更为明显或更不易被察觉的反应；是否有另一个更为有效的处理方案；以及这一实验是否涉及更广泛的社会理论。

接下来，我将通过韦斯利·舒尔茨（Wesley Schultz）及其同事在2007年针对社会规范与用电量之间的关系所开展的模拟实地实验来说明超越简单实验的价值所在。实验中，舒尔茨和同事选取了位于圣马科斯和加利福尼亚州的约300个家庭作为实验对象，并在他们家门前挂上了写有鼓励人们节约用电的信息的门挂牌。然后，舒尔茨和同事分别在一周后和三周后测量了这些信息对用电量的影响。更详细的实验设计描述参见图4.3。

实验是在两种不同的实验条件下开展的。在第一种实验条件下，实验对象收到的是一般的节能建议（例如使用风扇而不是空调）和他们各自的用电情况与所在小区的平均用电情况的对比信息。舒尔茨和同事称该实验条件为描述性规范实验条件，因为该小区的用电情况体现了一种典型行为（即描述性规范）。舒尔茨和同事在研究实施处理后该被测群体的用电情况时发现，无论从短期还是长期来看，这一处理似乎都没有任何效果。换言之，这一处理似乎并不能够“起作用”（图4.4）。

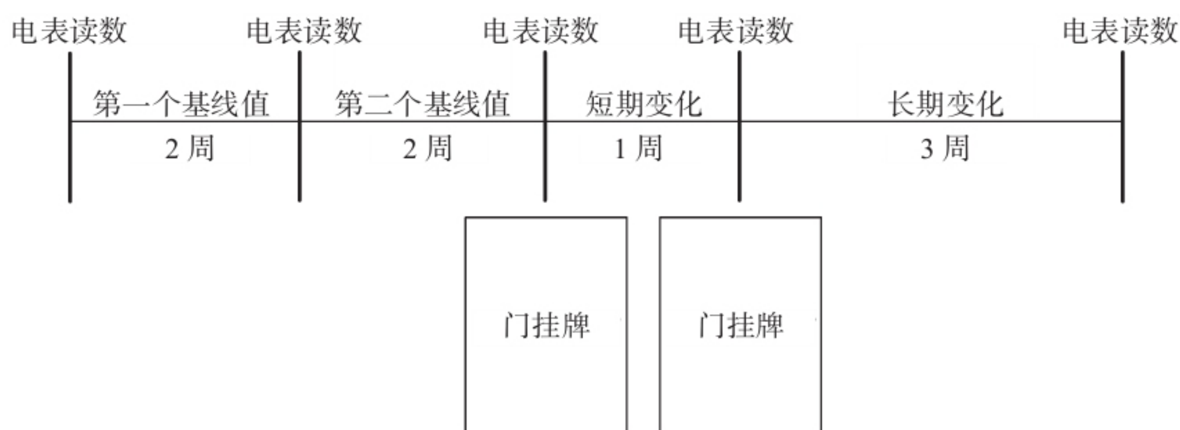


图4.3 舒尔茨等人的实验设计原理图。在该实地实验中，研究人员需在8周时间内拜访圣马科斯和加利福尼亚州约300个家庭。每次拜访时，研究人员都需手动记录下相应家庭电表的读数。在其中的两次拜访中，他们还需将写有家庭用电情况的门挂牌挂在相应家庭的门上。该项实验的研究课题是，这些信息的内容如何影响居民用电量。

幸运的是，舒尔茨和同事并没有满足于这一简单的分析。其实在实验开始前，他们曾推断，用电量大的居民，即高于平均用量的居民，其用电量可能会减少；而用电量小的居民，即低于平均用量的居民，其用电量则可能会增加。然后当他们仔细研究具体数据时发现，的确是这样的（图4.4）。因此，看起来没有任何效果的处理，实际上是产生了两个相互抵消的效果。其中在低用电量群体中所产生的适得其反的上涨效果则是“飞去来器效应”的一个例子，即某一处理产生了与原目标相反的效应。

第二种实验条件下的实验与第一种是同时进行的。在该实验条件下，被测家庭收到了几乎完全相同的门挂牌，上面写有一般节能建议和他们各自的用电情况与所在小区的平均用电情况的对比信息，不过还增加了一个小细节：对于低用电量居民，研究人员会额外附上表情符号☺；而对于高用电量居民，研究人员则会额外附上☹。这些表情符号旨在运用研究人员所称的指令性规范。指令性规范旨在表明什么是被人们赞同（或者不赞同）的行为，而描述性规范旨在表明人们应该做什么（Reno，Cialdini and Kallgren 1993）。

通过加上这个小小的表情符号，研究人员大大减轻了在低用电量居民中所出现的“飞去来器效应”（图4.4）。因此，这一简单的改变（受到一个抽象的社会心理学理论的启发）使原本似乎不可行的项目变得可行了，与此同时，这一实验还能使人们更进一步地理解社会规范是如何影响人类行为的。

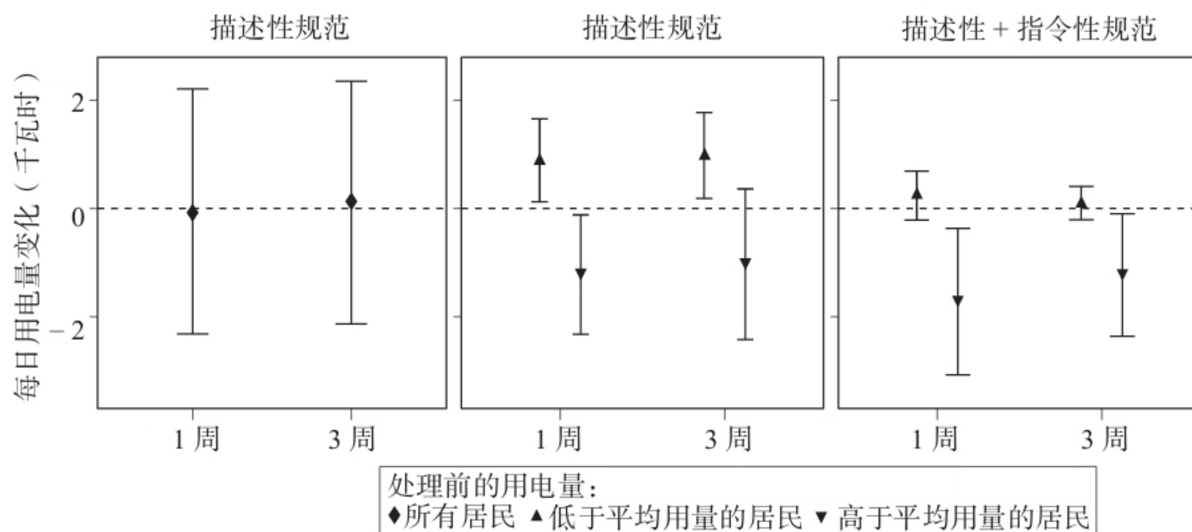


图4.4 舒尔茨等人的研究结果。图中第一栏表明，描述性规范所产生的平均处理效应（即对被测群体整体用电量的影响）几乎为零。但是，第二栏表明，该平均处理效应实际上包含了两种相互抵消的效应。对于高用电量居民，描述性规范会降低他们的用电量，但是对于低用电量居民，描述性规范反而会增加他们的用电量。最后，第三栏表明，第二种处理，即同时运用描述性和指令性规范，对高用电量居民产生的影响与第一种处理几乎相同，但是减轻了在低用电量居民中所出现的“飞去来器效应”。该图根据舒尔茨等人的研究结果绘制。

然而，你也可能会发现，舒尔茨及其同事的实验与其他实验有一点不同。那就是，这个实验并没有像随机对照实验那样有一个对照组。其实通过对比该实验设计跟雷斯蒂沃和范德里杰特的实验设计，就能阐明以下两种主要实验设计的差异。在“被试间”设计中，比如雷斯蒂沃和范德里杰特的实验设计，会设置一个实验组和一个对照组。但在“被试内”设计中，则是把参与者在接受实验处理前和后的行为做对比（Greenwald 1976; Charness, Gneezy, and Kuhn 2012）。在采用被试内设计的实验中，每个参与者似乎同时也是自己的对照组。被试间设计的优势在于降低了混淆变量的干扰（正如我在前文中提及的），而被试内设计的优势则在于提高了估计结果的准确性。然后，在后面的章节中我会针对如何设计数字实验给出建议，其中就会涉及混合设计，该设计融合了被试内设计的高准确度以及被试间设计的低干扰性（见图4.5）。

总的来说，舒尔茨及其同事的实验的设计和结果都说明了超越简单实验的价值。幸运的是，并非只有创造性的天才才能设计出这样的实验。社会科学家已经提出了三个理念来指导我们设计出更丰富的实验：（1）效度；（2）处理效应的异质性；（3）原理。也就是说，如果你在设计实验的时候牢记这三个理念，那么你自然而然地会设计出一个更有趣有用的实验。舒尔茨及其同事精妙的实验设计和令人兴奋的成果启发了一系列在部分程度上可以算作数字化的实地实验，我将通过对这些后续实验进行描述来进一步阐明应该如何应用这三个理念。然后你会发现，通过更为细致的设计、实施、分析和解读，你也能够实现从简单实验到丰富实验的突破。

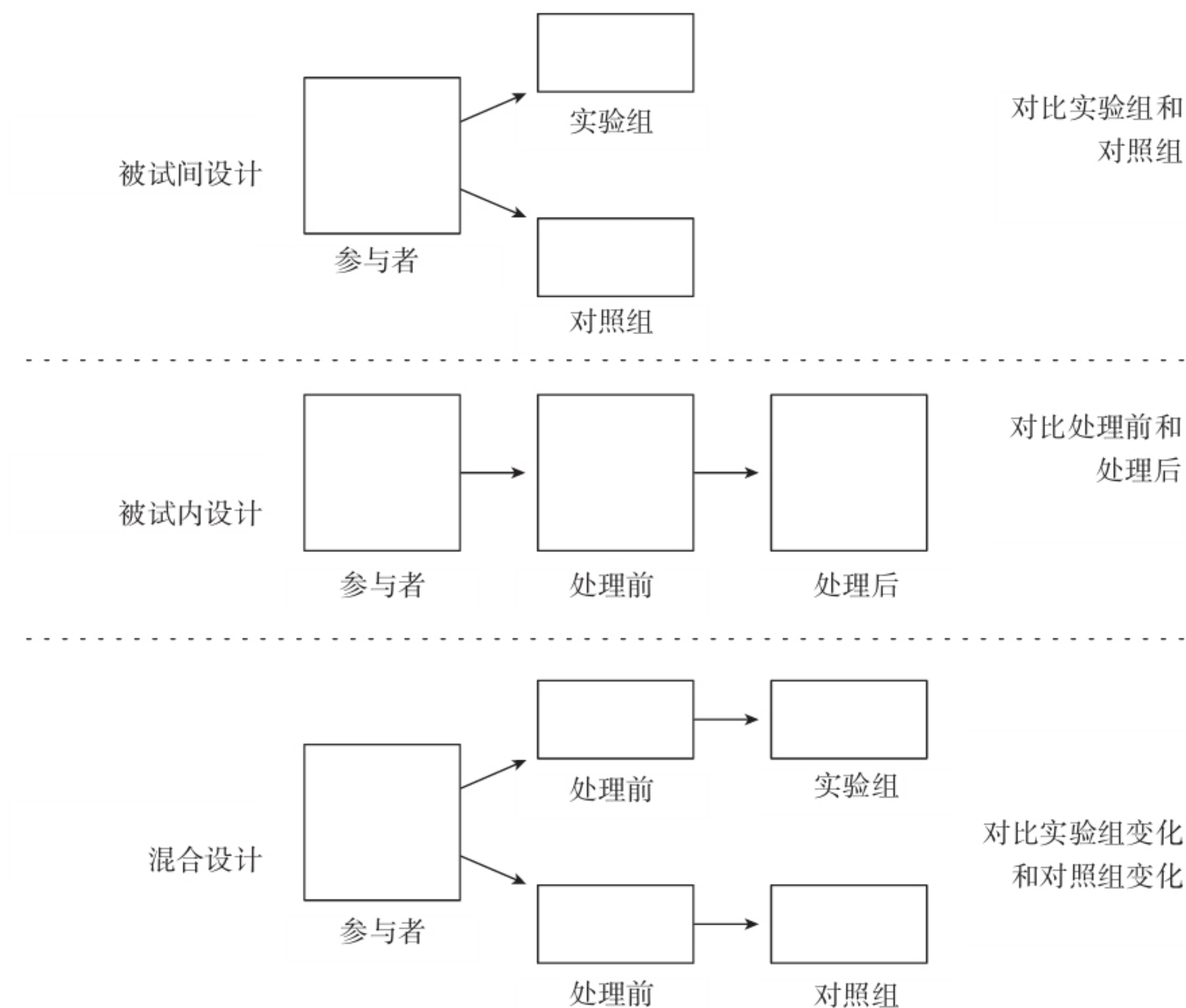


图4.5 上图为三种实验设计。首先，标准的随机对照实验采用的是被试间设计。采用被试间设计的实验的一个典型例子就是雷斯蒂沃和范德里杰特进行的实验，该实验旨在研究谷仓之星与用户对维基百科的贡献之间的关系：研究人员把参与者随机分为实验组和对照组，并给实验组的参与者每人一枚谷仓之星，然后比较两组的结果。第二种实验设计是被试内设计。舒尔茨及其同事进行的两个关于社会规范和用电量间关系的实验就用到了被试内设计：研究人员对比了参与者在接受处理前和接受处理后的用电量。被试内设计有利于提高统计数据的准确性，但是可能受到混淆变量的干扰（比如处理前和处理期间的天气变化）（Greenwald 1976； Charness, Gneezy, and Kuhn 2012）。被试内设计有时也被称为重复测量设计。最后是混合设计，它兼备了被试内设计的高准确度和被试间设计的低干扰性。在混合设计中，研究人员会比较实验组和对照组的结果变化。如果研究人员已经掌握了处理前信息（在许多数字实验中都是这样的），混合设计通常比被试间设计更可取，因为前者能够提高估计结果的精确度。

4.4.1 效度

效度是指实验结果符合某个一般性结论的程度。

没有一个实验是完美的，研究人员创造了大量的词汇来描述可能出现的问题。效度是指某一实验的结果符合某个一般性结论的程度。社会科学家发现可以将效度分成4种主要类型：统计结论效度、内部效度、构念效度和外部效度（Shadish, Cook, and Campbell 2001, chapter2）。掌握这些理念之后，你在评价和改进实验的设计和对实验进行分析时，心里就有谱儿了，而且它也会方便你和其他研究人员交流。

统计结论效度的核心在于对实验进行的统计分析是否正确。在舒尔茨及其同事的实验中，该效度的核心可能就是他们是否正确地计算了*P*值[注]。设计和分析实验所需的统计原则不在本书的内容范畴内，但数字时代的到来并没有让这些原则发生根本的改变。发生改变的是数字实验的数据环境，它创造了新的可能性，例如利用机器学习模型评估处理效应的异质性（Imai and Ratkovic 2013）。

内部效度的核心是实验步骤是否被正确地完成。在舒尔茨及其同事的实验中，该效度的核心可能就是随机分组、实施处理和测量结果。例如，你可能会担心研究助理的电表读数不准确。事实上，舒尔茨和同事也很担心这个问题，于是他们让助理把部分电表读了两次，幸运的是，两次的读数基本一致。总的来说，舒尔茨和同事的实验的内部效度似乎很高，但并不是所有实验都是这样：复杂的实地实验和在线实验在对正确的人实施正确的处理和测量每个人的结果方面，实际上经常会出现问题。幸运的是，数字时代有助于减少对内部效度的担忧，因为在数字时代，确保对参与者实施处理以及测量所有参与者的结果变得更加容易了。

构念效度的核心是数据和理论构念的匹配。正如第2章所讨论的，构念是社会科学家所论证的抽象概念。不幸的是，这些抽象概念并不

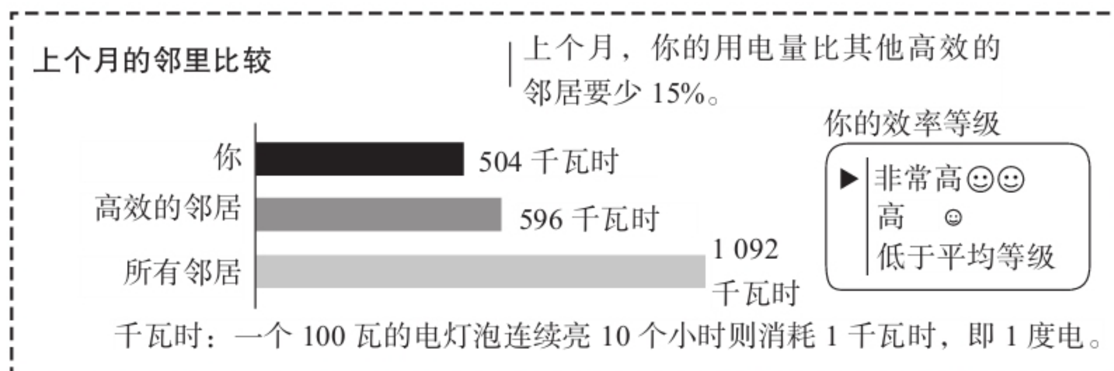
总是有明确的定义和度量。在舒尔茨等人的实验中，要想证明“指令性规范能够降低用电量”这一观点，需要研究人员设计一个能很好地代表指令性规范的处理方式（例如添加一个表情符号），并测量用电量。在模拟实验中，许多研究人员都是自行设计自己的处理方式并测量自己的结果。这一方法尽可能地确保了实验与所研究的抽象构念相匹配。在数字实验中，研究人员则通过与企业或政府合作来实施处理，并利用不间断运行的数据系统来测量结果，所以实验和理论构念之间的匹配可能没有那么紧密。因此，我认为相比于模拟实验，数字实验的构念效度会更令人担忧。

最后，外部效度的核心是该实验的结果能否被推广到其他情境中。就舒尔茨等人的实验而言，人们可能会问，如果在不同的情境中以不同的方式进行实验，那么给人们提供关于他们用电情况与所在小区的平均用电情况的对比信息和一个代表指令性规范的信号（例如一个表情符号）还能减少用电量吗？对于大多数精心设计、步骤正确的实验来说，对外部效度的担忧是最难解决的。在过去，有关外部效度的争论通常都只是一群人坐在一个房间里面，然后努力去想如果以不同的方式完成实验步骤，或是在一个不同的地方开展实验，或是有不同的参与者，那么会发生什么。幸运的是，数字时代使研究人员不用再进行这些缺乏数据支撑的猜测了，他们可以通过实证来评估外部效度。

因为舒尔茨等人的研究结果非常令人兴奋，所以欧电公司（Opower）与美国的公用事业单位联手对更多的住户实施了这一处理。根据舒尔茨等人的设计，欧电公司创建了包含两个主要模块的个性化家庭能源报告，其中一个模块显示某一住户相比于其邻居的用电量情况，同时会附上一个相应的表情符号，另一个模块则提供有关减少用电量的建议（图4.6）。然后，欧电公司与研究人员一起开展了随机对照实验，以评估这些家庭能源报告的影响。尽管这些实验的处理基本都是以非数字化的方式（通常是通过传统的信件邮寄）实施的，

但对结果的测量都是在物理世界使用数字设备进行的（例如使用电表）。此外，欧电公司在实验过程中还与电力公司进行了合作，使得研究人员能够直接访问电力公司的电力数据，而无须让研究助理挨家挨户去搜集这些信息。因此，欧电公司和合作伙伴以低可变成本的方式在大规模范围内成功地开展了这些半数字实地实验。

在最开始针对10个不同地点的60万户家庭进行的一组实验中，阿尔科特（Allcott）发现，家庭能源报告能够降低用电量。换句话说，这项规模更大的、涉及地理区域更多样化的研究所得出的结果与舒尔茨等人的结果在本质上是相似的。然后，在涉及101个不同地点的800万户家庭的后续研究中，阿尔科特再次发现了家庭能源报告能够降低用电量。但这一组更大规模的实验也揭示了一个有趣的新规律：在后来的实验中，家庭能源报告对用电量的影响减小了（图4.7），这一规律在任何单个的实验中都是无法被发现的。阿尔科特推测影响减小的原因是随着时间的推移，接受处理的参与者的类型不同了。更具体地说，客户环保意识越强的公用事业单位会越早参与这一项目，而且他们的客户也会更积极地响应。而随着客户环保意识较弱的公用事业单位加入，家庭能源报告的影响似乎就减小了。因此，正如实验中的随机分组能确保实验组和对照组是相似的一样，随机选择研究地点也能确保研究结论可以从一组参与者泛化至更普遍的总体（回想一下第3章有关抽样的内容）。也就是说，如果研究地点不是随机抽取的，那么即便是一个设计和实施都很完美的实验所得出的结论，在其泛化阶段也会遇到问题。



邻里比较模块

行动方案：根据你的用电情况和房屋概况为您选择的个性化方案		
速效对策 立刻就能做的事情	**巧妙购置** 花小钱省大钱	**超棒投资** 省大钱的大决定
调整电视的显示设置 新电视起初是为了让其在展厅呈现最佳的观看效果而设置的，但家用电视通常不需要这样的设置。 调整电视的显示设置最高能降低 50% 的耗电量，而且也不会影响画面质量。利用电视上的“显示”或“图片”菜单，调整“对比度”和“亮度”对能耗的影响最大。 调暗显示器还能延长电视的使用寿命。	**安装感应开关** 总是记不住关灯？感应开关能在你离开房间后自动将灯关掉，省心又省钱。 感应开关对频繁有人进出的房间（例如我们的家）和没有光线的地方（例如储物区域）来说是理想之选。 只需将标准的照明开关换成壁挂式感应开关即可，后者在大多数五金店都能买到。	**买一个新洗衣机** 用洗衣机洗衣服耗电量很大，尤其是用温水或热水洗。 事实上，当用温水或热水洗时，90% 的能源消耗都是给水加热所产生的。 一些节能洗衣机的用水量仅是老式洗衣机的一半，这就意味着它能帮你省钱。
每年每台最高能省 40 美元	每年最高能省 30 美元	每年最高能省 30 美元

行动方案模块

图4.6 家庭能源报告包括一个邻里比较模块和一个行动方案模块。经爱思唯尔（Elsevier）授权复制自Allcott（2011），图1和图2。

图4.7 111个测试家庭能源报告对用电量影响的实验结果。在后来加入该项目的地方，家庭能源报告对用电量的影响往往较小。阿尔科特认为，这一现象主要是因为用户环保意识越强的地方会越早加入这一项目。改编自Allcott（2015），图3。

上述111个实验共涉及来自美国各地约850万户家庭。这些实验均表明，家庭能源报告能够降低平均用电量，这与舒尔茨和同事最初从加利福尼亚州300户家庭那里得出的结论是一样的。除此以外，这些后续的实验还表明，家庭能源报告对用电量的影响力因地点而异。同时，这些实验也说明了有关半数字实地实验的两个更普遍的观点。首先，当开展实验的成本较低时，研究人员就能通过实证来解决外部效度相关的问题了。如果已经有一个不间断运行的数据系统正在对实验结果进行测量的话，那么就满足了这一条件。因此，研究人员应该留意那些已经在被记录的有趣且重要的行为，然后在现有的测量基础设施上设计实验。其次，这些实验提醒我们，数字实地实验并非只能在线进行，我认为它们会越来越普及，许多结果都能通过建筑环境中的传感器来测量。

统计结论效度、内部效度、构念效度和外部效度这4种效度为研究人员提供了一个思考依据，有助于他们评定某一特定实验的结果是否能够证明一个更为普遍的结论。相比于模拟时代的实验，数字时代的实验应该更易于通过实证来解决外部效度的问题，同时也更易于确保内部效度，而构念效度的问题则可能变得更具挑战性，尤其是在那些和企业合作的数字实地实验中。

4.4.2 处理效应的异质性

实验通常测量的是平均效应，但它对每个人产生的效应很可能是不一样的。

超越简单实验的第二个关键概念是处理效应的异质性。舒尔茨等人的实验有力地说明了同样的处理对不同类型的人会产生不同的效应（图4.4）。但在大多数模拟实验中，研究人员关注的是平均处理效应，因为实验只有少量参与者，而且研究人员对他们也知之甚少。而

在数字实验中，通常会有更多的参与者，且研究人员对他们的了解也更多。在这样一个不同的数据环境中，继续只关注平均处理效应的研究人员就会错失三方面的信息：一个处理是如何起作用的、如何对其进行改进以及如何才能让最有可能受益的人接受处理。因为这是要靠评估处理效应的异质性才能获得的信息。

以下两个针对家庭能源报告的额外研究可以被看作处理效应异质性的两个示例。在其中一项研究中，阿尔科特按照实施处理前的用电量将60万户家庭进一步划分为10个等级，然后分别评估了家庭能源报告对它们的影响。舒尔茨等人发现了家庭能源报告对高用电量居民和低用电量居民的影响存在着差异，而阿尔科特则发现，单就高用电量居民或低用电量居民来说，其内部对家庭能源报告的反应也存在着差异。例如，用电量最高的居民（即10个等级中最靠上的居民）节约的电量是用电量居中（就所有高用电量居民来说）的居民的两倍（图4.8）。此外，阿尔科特在这一研究中还发现，并不存在“飞去来器效应”，即使在用电量最低的居民中也没出现这一效应（图4.8）。

图4.8 阿尔科特的研究中呈现出的处理效应的异质性。不同等级的用户节约的电量也不同。改编自Allcott（2011），图8。

在另一项研究中，科斯塔（Costa）和卡恩（Kahn）猜测，家庭能源报告的有效性可能会因参与者的政治意识形态不同而有所差异，某些意识形态的参与者的用电量可能还会因这一处理而增加。换句话说，他们猜测家庭能源报告对某些类型的人可能会产生“飞去来器效应”。为了验证他们的猜测，科斯塔和卡恩将欧电公司的数据与从一个第三方的数据整合者那里购买的数据进行了整合，其中后者的数据包括政党登记、给环境组织的捐赠以及可再生能源在家庭生活中的使用等方面的信息。通过整合后的数据集，科斯塔和卡恩发现，家庭能源报告对政治意识形态不同的参与者所产生的影响大致是相似的，而且各组均未表现出“飞去来器效应”的迹象（图4.9）。

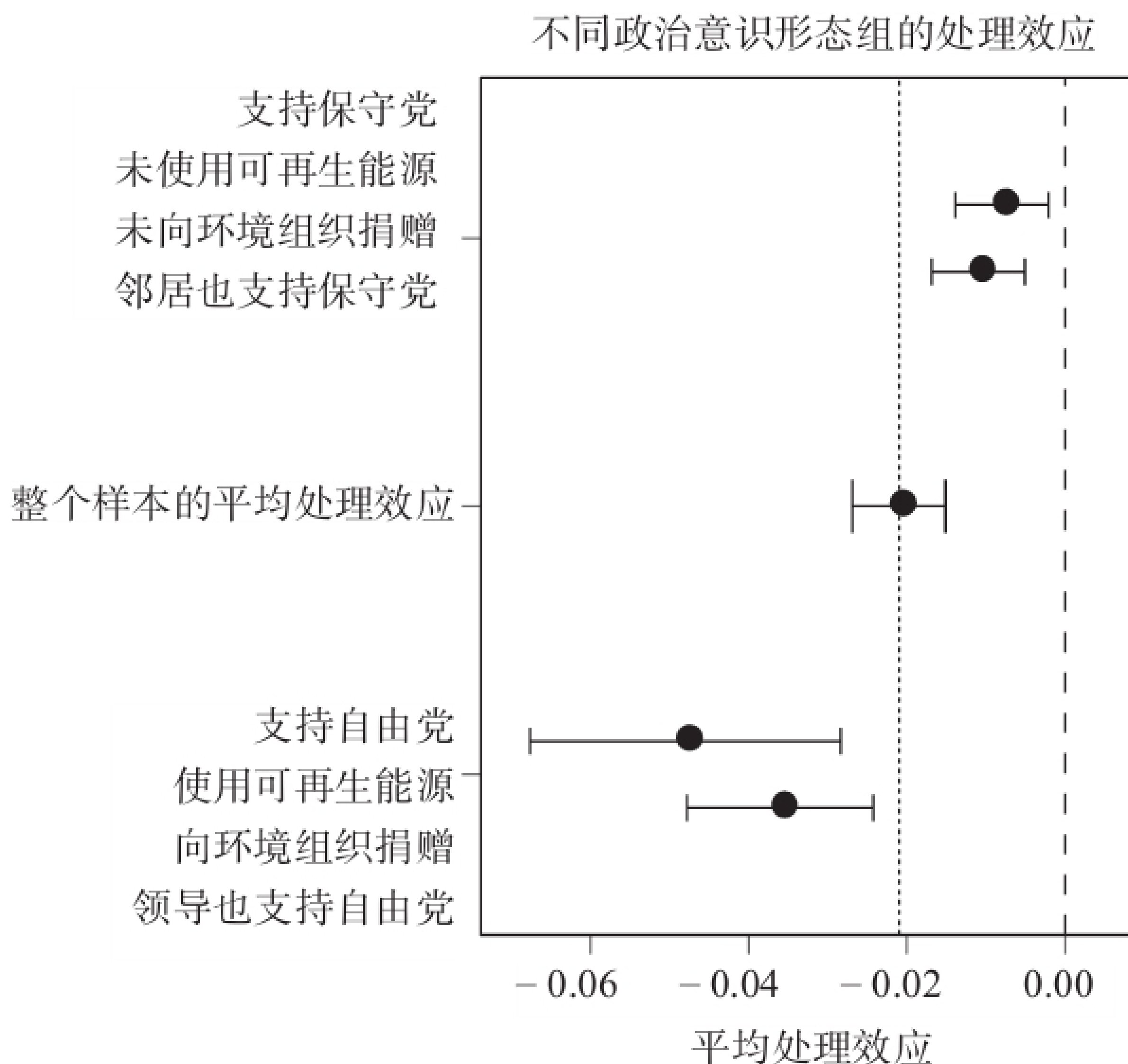


图4.9 科斯塔和卡恩的研究中呈现出来的处理效应的异质性。科斯塔和卡恩对整个样本产生的平均处理效应的估值是－2.1%（－1.5%，－2.7%）。将通过实验得到的信息与家庭信息合并后，他们利用一系列统计模型评估了对特定群组的处理效应。每个群组都有两个估值，因为这些估值取决于他们统计模型中所包含的协变量。正如该实验所示，处理效应可能会因人而异，而利用统计模型得出的处理效应的估值也取决于这些模型的细节（Grimmer，Messing，and Westwood 2014）。改编自Costa and Kahn（2013），表3和表4。

正如这两个研究所示，数字时代让我们实现了从评估平均处理效应到评估处理效应的异质性的过渡，因为我们能拥有更多的参与者，而且对他们有更多的了解。了解处理效应的异质性能让研究人员为某

一处理找到最有效的对象，能提供促进新理论发展的事实，还能为可能的原理提供线索，我接下来就将介绍原理。

4.4.3 原理

实验测量发生了什么，原理则解释这是为什么以及如何发生的。

超越简单实验的第三个关键概念是原理。原理能告诉我们一个处理为什么能产生影响或如何产生影响。弄清楚原理的过程有时也被称作寻找中介变量。尽管通过实验可以很好地评估因果效应，但实验设计的初衷往往并不是为了揭示原理。数字实验能以两种方式帮助我们找出原理：（1）使我们能够搜集更多的过程数据，（2）使我们能够测试许多相关的处理。

因为很难正式定义原理（Hedström and Ylikoski 2010），所以我将先从一个简单的示例开始，即青柠和维生素C缺乏病的关系（Gerber and Green 2012）。在18世纪，医生普遍都知道，如果水手们吃青柠的话，他们就不会得维生素C缺乏病。维生素C缺乏病是一种可怕的疾病，所以这是一个非常有用的信息。但医生并不知道青柠为什么能够预防它。直到将近200年后的1932年，科学家才证明了青柠之所以能预防该病是因为含有维生素C（Carpenter 1988）。也就是说，在该示例中，维生素C就是青柠能预防这种病的原理（图4.10）。找出原理具有很重要的科学意义，许多科学都是关于理解事情为什么会发生的。它同时也具有很重要的实践意义，一旦我们理解了一个处理起作用的原理，就有可能开发出效果更好的新处理方案。

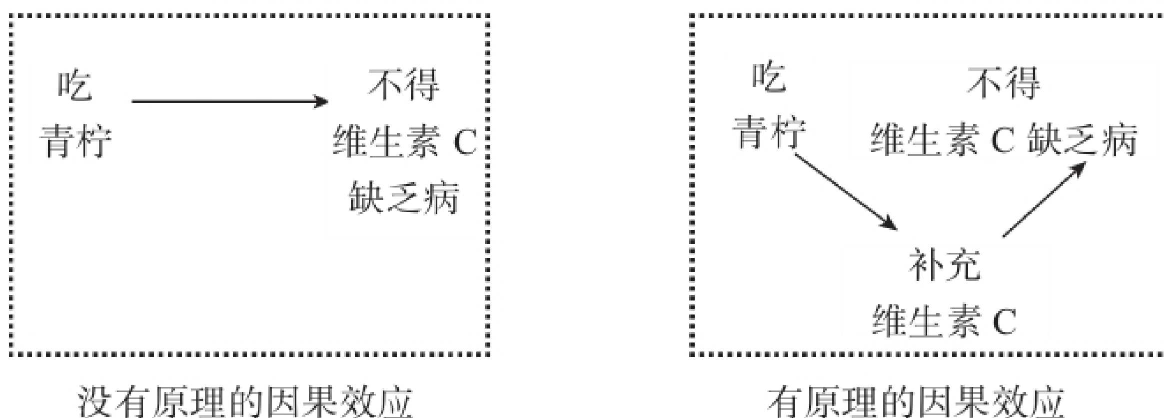


图4.10 青柠能够预防维生素C缺乏病，其原理是它含有丰富的维生素C。

不幸的是，找出原理是非常困难的。不同于青柠和维生素C缺乏病，在许多社会情境中，处理很可能是通过多个相互关联的途径产生影响的。但在前面所描述的有关社会规范和用电量的研究中，研究人员就试图通过搜集过程数据并测试相关处理来找出原理。

找出可能原理的一种方法是搜集某一处理如何影响可能原理的过程数据。例如，阿尔科特曾指出，家庭能源报告能够让人们减少用电量。但这些报告是如何减少用电量的呢？原理是什么呢？在一项后续研究中，阿尔科特和罗杰斯与一家电力公司进行了合作。该电力公司通过一项回赠活动了解了有哪些用户将其家用电器升级为更节能的产品。通过这些信息，阿尔科特和罗杰斯发现，收到家庭能源报告且升级了家电的家庭数量仅比未收到报告却也升级了家电的家庭数量稍微多一些，因此升级家电所节约的电量仅占收到报告的家庭所节约的总电量的2%。换句话说，升级家电并不是家庭能源报告降低用电量的主要原理。

找出原理的第二种方法是在实验中设置彼此稍有不同的处理。例如，在舒尔茨等人的实验以及后续所有关于家庭能源报告的实验中，参与者收到的家庭能源报告（即实验处理）主要包括两部分：（1）节能建议，（2）其与邻居的用电量对比信息（图4.6）。因此，引起变

化的原因可能是节能建议，而不是与邻居的对比信息。为了验证这一可能性，费拉罗（Ferraro）、米兰达（Miranda）和普赖斯（Price）与佐治亚州亚特兰大附近的一家水务公司合作开展了一项有关节约用水的实验。该实验涉及了约100000户家庭，他们将这些家庭分成了以下4组：

- 收到节水建议的一组；
- 收到节水建议以及一条呼吁节水的倡议的一组；
- 收到节水建议、一条呼吁节水的倡议以及其与邻居在用水量方面的对比信息的一组；
- 未收到任何信息的对照组。

研究人员发现，只有节水建议的处理在短期（一年）、中期（两年）和长期（三年）内对参与者的用水量都没有影响；节水建议外加节水倡议的处理能让参与者减少用水量，但仅限于短期内；节水建议、节水倡议以及邻里对比信息的处理则在短期、中期和长期内都能让参与者减少用水量（图4.11）。采取分类处理的这类实验是找出处理的哪一部分或哪些部分是造成影响的原因的一种好方法（Gerber and Green 2012, section 10.6）。例如，费拉罗和同事的实验就表明了只有节水建议是不能让参与者减少用水量的。

理想的情况是，我们能从上述类型的处理分层设计上升到完全析因设计，有时也称为2^k析因设计。在该设计中，研究人员会对三个因素所有可能的组合都进行测试（表4.1），这样便能对每个因素单独的效应和组合起来的效应进行全面的评估。例如，费拉罗和同事的实验就没有表明只有邻里对比信息是否能带来行为的长期改变。在过去，开展完全析因实验是很困难的，因为需要大量的参与者，而且需要研

究人员能够精确地控制和实施大量的处理。但在某些情形下，数字时代消除了这些组织实施上的限制。

图4.11 费拉罗、米兰达和普赖斯的实验结果。研究人员于2007年5月21日将不同内容的处理信息发出，然后分别于2007年、2008年和2009年的夏天测量了结果。通过采取不同类型的处理，研究人员希望能对原理有更好的理解。竖线代表的是研究人员估计的置信区间。实际研究材料可参见Bernedo，Ferraro，and Price（2014）。改编自Ferraro，Miranda，and Price（2011），表1。

表4.1 三个因素的完全析因设计中的处理示例
三个因素分别是建议、倡议和邻里信息

处理	特征
1	无处理
2	建议
3	倡议
4	邻里信息
5	建议＋倡议
6	建议＋邻里信息
7	倡议＋邻里信息
8	建议＋倡议＋邻里信息

综上所述，原理，即一项处理产生效应的途径，是极其重要的。数字时代的实验有助于研究人员通过搜集过程数据和采取完全析因设计了解原理。然后研究人员可以利用专门用于测试原理的实验，直接验证通过这些方法所了解到的原理。

总的来说，效度、处理效应的异质性和原理这三个概念为设计和评价实验提供了一个有力的框架。这些概念有助于研究人员超越只针

对“什么能够起作用”的简单实验，进而设计出与理论联系更紧密的、能够揭示处理在何种情境下以及为什么会产生效应的丰富实验，它们甚至能帮助研究人员设计出更有效的处理方案。在这一有关实验的概念背景下，接下来我将介绍如何才能开展实验。

---

1. P值就是当原假设为真时所得到的样本观察结果或更极端的结果出现的概率。——编者注

4.5 使实验成为现实

即便你不在一家大型科技公司工作，也能开展数字实验。你可以自行开展或与能帮助你的人（以及你能帮助的人）合作开展。

在这一点上，我希望大家对自己有望开展数字实验感到兴奋。如果你在一家大型科技公司工作，那么这类实验对你来说可能已经是家常便饭了。但如果你不在一家科技公司工作，可能就会认为自己无法开展这类实验。幸运的是，事实并非如此。只需一点创造力和努力，每个人都能开展数字实验。

首先，区分自行开展和与有能力的人合作开展这两种主要方式是很有帮助的。而且自行开展实验也有几种不同的方式：利用现有环境开展实验、创建自己的实验或是创建自己的产品以进行反复实验。通过下面的例子你可以看出，这些方法中并不存在对所有情形都最适用的方法，最好是将它们看作在成本、控制、真实和道德伦理这4个主要维度上各有利弊的方法（图4.12）。

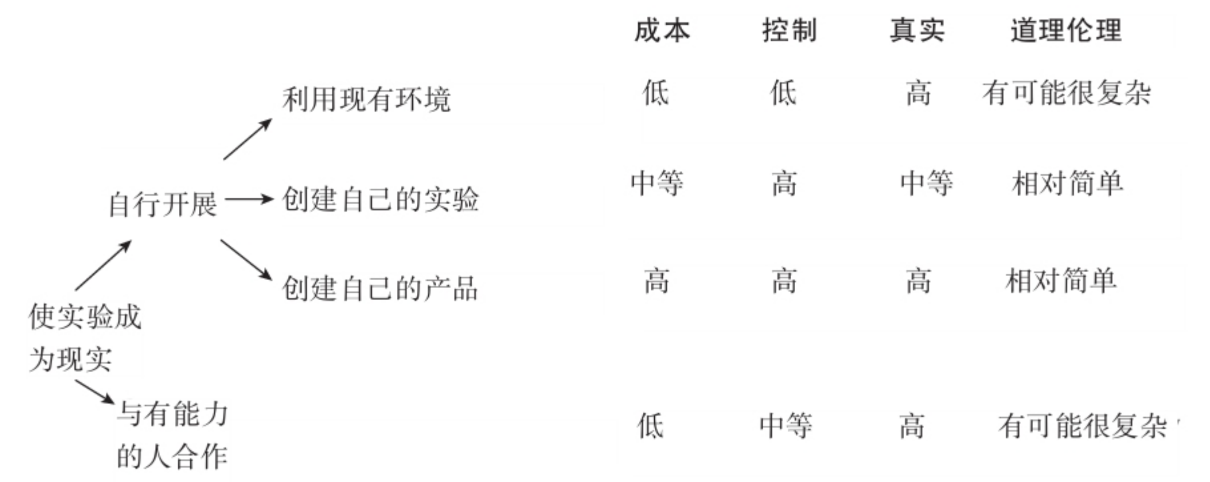


图4.12 各种开展实验的方式的利弊总结。成本，是指研究人员所花费的时间和金钱。控制，是指在招募参与者、随机分组、实施处理和测量结果方面做你想做的事

情的能力。真实，是指实验环境与日常生活中所遇到的情形相匹配的程度；需注意的是，高匹配程度对验证理论来说并不总是重要的（Falk and Heckman 2009）。道德伦理，是指拥有好的出发点的研究人员应对可能出现的道德伦理挑战的能力。

4.5.1 利用现有环境开展实验

你可以在现有环境中开展实验，这通常无须进行任何编码或与别人合作。

从逻辑上讲，开展一项数字实验，最简单的方法就是在现有环境的基础上开展实验。采取这一方式可以开展相当大规模的实验，且无须与企业合作或是进行大量的软件开发。

例如，珍妮弗·多里亚克（Jennifer Doleac）和卢克·斯泰因（Luke Stein）就曾利用一个类似于克雷格列表网（Craigslist）的在线商城开展了一项旨在测量种族歧视的实验。在实验中，他们为上千个音乐播放器做广告宣传，然后通过系统地改变卖家的特征研究了种族对经济交易的影响。此外，他们还借助实验的规模评估了什么时候该影响会更大（处理效应的异质性），并对该影响产生的可能原因（原理）给予了一些解释。

多里亚克和斯泰因的音乐播放器广告主要在三个方面存在着差异。首先是卖家的特征不同，表现在照片中拿音乐播放器的手的不同（肤色是白的、黑的、白的且有文身，见图4.13）。其次是要价不同（90美元、110美元、130美元）。最后是广告内容的质量的不同（高质量和低质量，例如单词是否有大小写和拼写方面的错误）。因此，广告采用了3×3×2的设计，并被投放到了从小城市（例如印第安纳州的科科莫和内布拉斯加州的北普拉特）到大都市（例如纽约和洛杉矶）300多个地方的当地市场中。

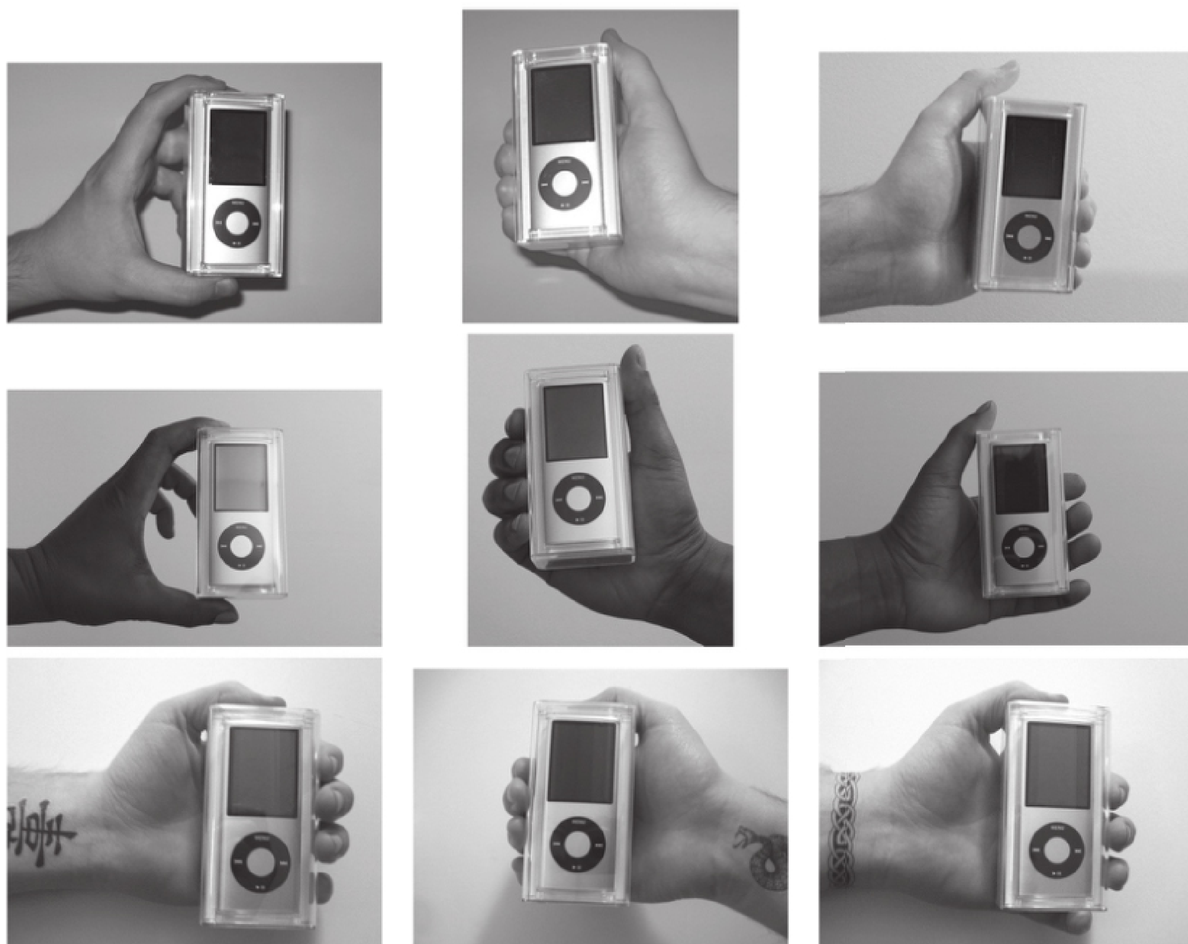


图4.13 多里亚克和斯泰因实验中手的照片。为了衡量在线商城中存在的种族歧视，音乐播放器分别由具有不同特征的卖家出售。经约翰·威利父子出版公司（John Wiley and Sons）许可复制自Doleac and Stein（2013），图1。

总的来说，白人卖家的销售情况比黑人卖家的要好，有文身的卖家的销售情况则居中。例如，白人卖家的音乐播放器有更多的买家，且其最终的成交价也较高。除了这些平均效应以外，多里亚克和斯泰因还评估了效应的异质性。例如，早期理论的一个预测是，在买家竞争越激烈的市场，歧视会越少。通过将市场中买家出价的次数作为衡量买家竞争程度的指标，研究人员发现，在买家竞争程度较低的市场，黑人卖家收到的出价确实较少也较低。此外，通过对比高质量广告和低质量广告的播放器销售情况，多里亚克和斯泰因发现，广告质量对黑人卖家和有文身的卖家所受到的歧视没有影响。最后，借助广

告投放范围包含300多个市场这一优势，研究人员发现，黑人卖家在犯罪率和居住隔离程度均偏高的城市受到的歧视也会更多。这些结果均无法准确地解释为什么黑人卖家的销售情况会较差，但是，如果将这些结果与其他研究结果相结合，研究人员便能对不同类型的经济交易中种族歧视的原因给出一些理论上的解释了。

另外一个表明研究人员可以利用现有系统开展数字实地实验的例子是阿尔努·范德里杰特和同事在2014年针对成功的关键的研究。在生活的许多方面看似相似的人，其最终的发展结果会截然不同。对此，一个可能的解释是，一个人所具有的小且基本随机的优势会一直伴随着他并随着时间的推移而增长，这一过程被研究人员称为优势累积。为了确定起初微小的成功会一直存在还是会逐渐消失，范德里杰特和同事对4个不同的系统进行了干预，即让随机选中的参与者获得相应的成功，然后测量这些随机分配的成功的后续影响。

更具体地说，他们对以下4个系统进行了干预：（1）在众筹网站Kickstarter上，范德里杰特和同事投资了随机选择的项目；（2）在产品点评网站Epinions上，他们正面评价了随机选择的评论；（3）在维基百科上，他们奖励了随机选择的贡献者；（4）在请愿网站Change.org上，他们在随机选择的请愿书上签下了自己的名字。然后他们在4个系统中均发现了极为相似的结果：与原本毫无区别的同一个系统的其他用户相比，在起始阶段被随机选中并获得相应成功的参与者，其后续的成功也会更多（图4.14）。许多系统中都存在着这样的规律，这便增加了上述结果的外部效度，因为这一事实降低了这个规律只是某一特定系统产物的可能性。

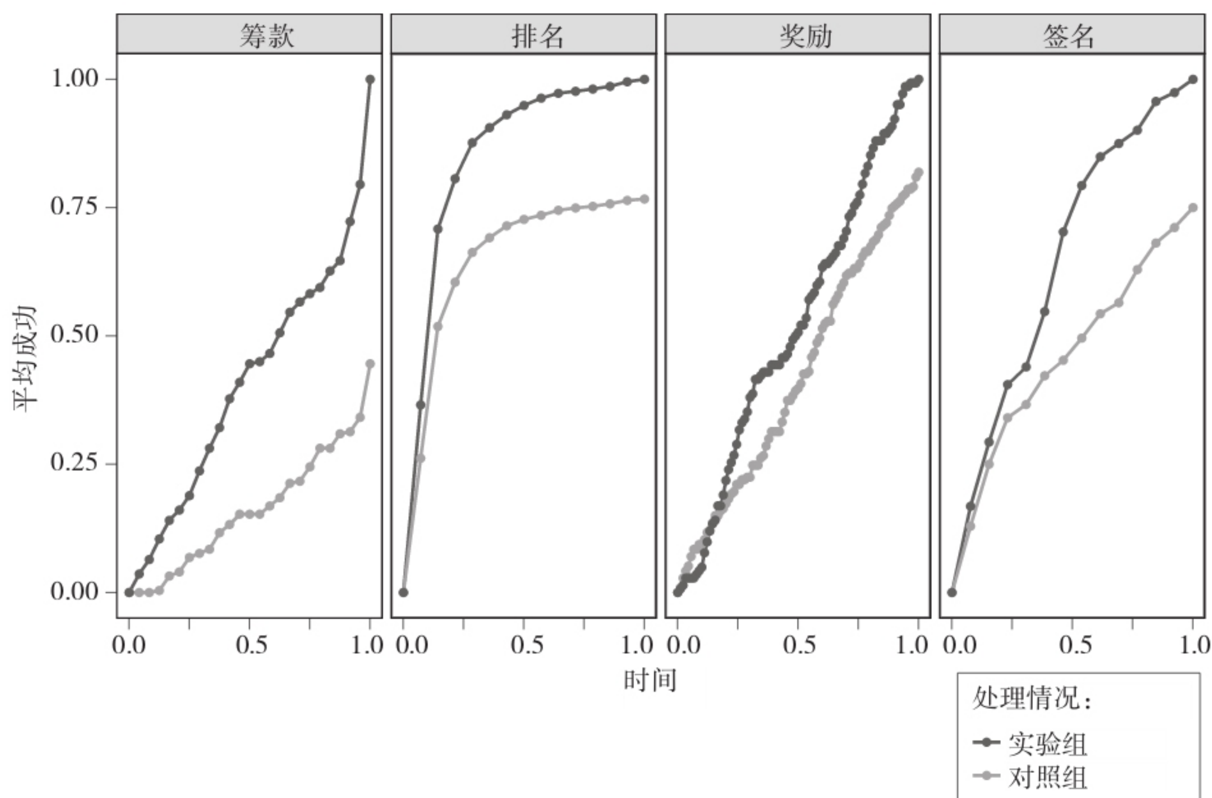


图4.14 4个不同系统中随机分配的成功的长期效应。(1)在众筹网站Kickstarter上，范德里杰特和同事投资了随机选择的项目；(2)在产品点评网站Epinions上，他们正面评价了随机选择的评论；(3)在维基百科上，他们奖励了随机选择的贡献者；(4)在请愿网站change.org上，他们在随机选择的请愿书上签下了自己的名字。改编自van de Rijt et al.(2014)，图2。

这两个例子表明，研究人员在不与企业合作或不构建复杂数字系统的情况下，也能开展数字实地实验。此外，表4.2列出了更多的实验，用以说明研究人员在利用现有系统的基础设施来实施处理、测量结果方面有着哪些可能性。对研究人员来说，这些实验的成本相对较低，且较贴近真实生活，但其对参与者、处理和需要测量的结果的掌控程度有限。此外，对于仅通过一个系统开展的实验，研究人员需要注意，实验的处理效应可能是在该系统特有的机制（例如Kickstarter对项目的排序方式或Change.org对请愿书的排序方式，更多内容可参见第2章的算法干扰）的推动下产生的。最后，当研究人员利用正在运行的系统开展实验时，会遇到一些棘手的道德伦理问题，有可能会对参与者、非参与者和系统造成伤害。第6章将更详细地探讨这些道德伦

理问题，同时范德里杰特的研究附录对此也进行了很好的讨论。但这些利用现有系统开展实验的权衡取舍并不适用于所有项目，因此正如我即将介绍的，有些研究人员会创建自己的实验系统。

表4.2 在现有系统中开展实验的示例

研究主题	参考文献
谷仓之星对人们编辑维基百科的贡献的影响	Restivo and van de Rijt（2012, 2014）；van de Rijt et al.（2014）
反骚扰信息对种族主义推文的影响	Munger（2016）
拍卖方式对成交价的影响	Lucking-Reiley（1999）
声誉对在线拍卖价格的影响	Resnick et al.（2006）
卖家的种族对其在易贝上拍卖棒球卡的影响	Ayres, Banaji, and Jolls（2015）
卖家的种族对其出售的音乐播放器的销量的影响	Doleac and Stein（2013）
爱彼迎（Airbnb）上租客的种族对其租金的影响	Edelman, Luca, and Svirsky（2016）
众筹网站 Kickstarter 上的捐款对项目成功的影响	van de Rijt et al.（2014）
所属种族对住房租金的影响	Hogan and Berry（2011）
产品点评网 Epinions 上的好评对日后评价的影响	van de Rijt et al.（2014）
签名对请愿成功的影响	Vaillant et al.（2015）；van de Rijt et al.（2014）；van de Rijt et al.（2016）

4.5.2 创建自己的实验

创建自己的实验可能需要很高的成本，但能使你开展自己想要的实验。

除了利用现有环境开展实验以外，大家还可以创建自己的实验。该方法主要的优势是可控性，也就是说，如果你自行创建实验的话，就可以创建自己想要的环境和处理。这些定制的实验环境可以为测试那些在自然环境中无法测试的理论创造机会。但创建自己的实验也有弊端，最主要的就是成本可能会很高，而且你所创建的环境可能无法具有自然存在的系统的真实性。此外，创建自己的实验的研究人员还必须有招募参与者的策略。在利用现有系统开展实验时，研究人员基本上是将实验通过系统顺便带给了参与者。但如果是创建自己的实验，研究人员则需要招募参与者参加实验。幸运的是，像机器人MTurk这样的服务平台为研究人员招募实验参与者提供了便利的渠道。

可被用来说明定制环境在测试抽象理论方面优势的例子是格雷戈里·休伯（Gregory Huber）、塞思·希尔（Seth Hill）和加布里埃尔·伦兹（Gabriel Lenz）2012年的数字实验室实验。该实验探究了民主治理可能存在的实际限制。早期针对实际选举的非实验性研究表明，选民无法对现任执政者的表现进行准确的评估。具体来说，选民似乎会因以下三个因素而在评估时出现偏差：（1）他们关注的是现任执政者近期而非一直以来的表现；（2）他们可能会被华而不实的言论、诬陷和营销信息所操纵；（3）他们可能会被与现任执政者政绩无关的事件所影响，例如当地运动队的获胜或天气。但在这些早期的研究中，研究人员很难将上述任何一个因素与真实复杂的选举中的其他事情隔离开来。因此，休伯和同事创建了一个高度简化的投票环境，以将上述三种偏差分别隔离出来，然后对其进行实验研究。

下述实验设置听起来很不真实，但大家需要记住的是，真实并不是实验室实验的一个目标。实验室实验的目标是将你试图研究的过程明确地隔离出来，在更真实的研究中，这一目标有时反倒难以实现了（Falk and Heckman 2009）。此外，在休伯和同事的研究中，他们认为，如果选民在其创建的高度简化的环境中无法有效评估执政者的政绩，那么他们在更真实、更复杂的环境中就更没办法有效评估了。

休伯和同事是通过机器人MTurk招募参与者的。只要参与者签署了知情同意书并通过一个简短测试，就会被告知他正在参与一项共有32轮的游戏，通过游戏就能赢取可兑换现金的代币。在游戏开始时，每个参与者会被告知他有一个分配器，该分配器会在每轮游戏中免费给他发放代币，并告诉他有的分配器发放的代币要多于其他分配器。此外，每个参与者还会被告知，在16轮游戏后，他将有机会选择是继续保留现有的分配器还是要求重新分配一个分配器。鉴于大家已经了解了休伯和同事的研究目标，所以你们应该明白了，这里的分配器代表的就是一个政府，16轮游戏后的选择代表的就是选举，但参与者并未意识到研究的目标。休伯和同事共招募了约4000名参与者，每名参与者在完成这项大约需8分钟的任务后会获得1.25美元的酬劳。

正如前面提到的，早期研究的一个发现是，选民会因执政者完全无法掌控的事情，例如当地运动队的胜利或天气，对其政绩做出过高或过低的评估。为了评估参与者的选择是否会受到其所在环境中完全随机的事件的影响，休伯和同事在实验中增加了抽奖环节，即在第8轮或第16轮游戏时，参与者会进行一次随机抽奖，其中有些人会赢5000分，有些人赢0分，有些人则输5000分。休伯和同事旨在用这一抽奖环节来模拟那些与执政者政绩无关的好的或坏的消息。尽管参与者被明确告知，抽奖与他们的分配器表现无关，但抽奖结果还是会影响参与者的选择。在抽奖中赢了5000分的参与者更有可能保留其分配器，而且相比于将抽奖设置在第8轮，将其设置在第16轮时（刚好在选择是否更换分配器之前）这一影响会更加明显（图4.15）。根据这些结果和其论文中其他几项实验的结果，休伯和同事得出结论，即便在简化的环境中，选民也难以做出明智的决定，这一结论影响了日后关于选民决策的其他研究（Healy and Malhotra 2013）。休伯和同事的实验表明，机器人MTurk可被用来为旨在准确测试非常具体的理论的实验室实验招募参与者。这个实验同时也说明了创建自己的实验环境的价值：很难想象在其他环境中如何将这些因素如此明确地隔离出来。

图4.15 休伯、希尔和伦兹的实验结果。在抽奖中赢了5000分的参与者更有可能保留其分配器，而且相比于将抽奖设置在第8轮，将其设置在第16轮时这一影响会更加明显。改编自Huber，Hill，and Lenz（2012），图5。

除了创建类似实验室的实验环境，研究人员还能创建更贴近现实的实验环境。例如，森托拉（Centola）就构建了一个数字实地环境，以研究社交网络结构对行为传播的影响。他研究的问题需要他观察同一行为在多个群体内（这些群体仅在社交网络结构方面存在着差异，其他方面基本一致）的传播。要想做到这一点，唯一的方法就是创建一个定制实验。在这种情况下，森托拉创建了一个基于网络的健康社区。

森托拉在健康网站上发布广告，招募了约1500名参与者。当参与者进入被称为“健康生活方式网”的在线社区时，他们需签署知情同意书，然后森托拉会为其分配“健康伙伴”。森托拉分配健康伙伴的方式使他能够在不同群组内构建不同的社交网络结构：有些群组是随机网络（即每个人被选为健康伙伴的概率是一样的），其他群组则是集群网络（即有些人被选为健康伙伴的概率要大一些）。然后，森托拉在每个网络中引入了一种新的行为：注册一个拥有额外健康信息的新网站的机会。每当有人注册这个新网站时，他所有的健康伙伴都会收到关于他注册行为的一封电子邮件。森托拉发现，相比于随机网络，注册新网站的行为在集群网络中传播得更远、更快。这一发现与现有的一些理论相悖。

总的来说，创建自己的实验能让你有更多的掌控力，能让你构建出最有利于隔离研究对象的环境。我很难想象上述两个实验在现有的环境中该如何开展。此外，创建自己的系统还能减少利用现有环境开展实验所面临的道德伦理问题。但创建自己的实验也会遇到实验室实验所面临的许多问题，比如招募参与者和对真实性的担心。尽管实验可以在相对简单的环境中进行（例如休伯、希尔和伦兹针对选举的研究），也可以在相对复杂的环境中进行（例如森托拉针对网络和传播

的研究），但创建自己的实验还有最后一个缺点，那就是可能会既费钱又费时。

4.5.3 创建自己的产品

创建自己的产品是一种高风险、高回报的方法。但如果成功了，就能带来正反馈循环，使你能够开展与众不同的研究。

有些研究人员会在创建自己的实验这一方法上更进一步，进而创建自己的产品。这些产品会吸引用户，然后便成了开展实验和其他类型研究的平台。例如，明尼苏达大学的一组研究人员就创建了MovieLens（意为“电影镜头”），一个免费的、非商业性质的、个性化的电影推荐网站。自1997年以来，MovieLens一直在运营，在此期间，网站搜集了25万名注册用户针对3万多部电影的2000多万条评分数据（Harper and Konstan 2015）。研究人员利用MovieLens的活跃用户群体开展了一系列精彩的研究，从测试有关公共产品捐款的社会科学理论到处理推荐系统所面临的算法挑战。（有关这些研究的全面性的回顾，可参阅Harper and Konstan 2015）。如果研究人员没有这样一个他们能够完全掌控的实际运行的产品，那么这些实验中的许多实验都将无法开展。

不幸的是，创建自己的产品是非常困难的，这就像是在创办一家公司一样：高风险、高回报。如果成功了，就对实验的掌控力来说，利用这一方法所开展的实验和创建自己的实验基本是一样的；就真实性和招募参与者来说，利用这一方法开展实验和利用现有环境开展实验是一样的。此外，这一方法可能会带来正反馈循环，即更多的研究会让产品变得更好，更好的产品会吸引更多的用户，有了更多的用户便可以开展更多的研究，以此类推（图4.16）。换句话说，一旦正反馈循环开始了，研究就会变得越来越容易。尽管这个方法目前来说非

常困难，但我希望随着技术的改进，它将变得更加实用。然而在那之前，如果研究人员想要利用一个产品，更直接的策略是与一个公司合作，这也是我接下来要讲的话题。

图4.16 如果你能成功创建自己的产品，便能受益于正反馈循环：研究能让产品变得更好，这样便能吸引更多的用户，有了更多的用户便能开展更多的研究。这类型的正反馈循环是非常难实现的，但它能让研究人员开展之前不可能的实验。MovieLens就是一个成功创造了正反馈循环的研究事例（Harper and Konstan 2015）。

4.5.4 与有能力的组织合作

合作可以降低成本、扩大规模，但也可能会改变你所能使用的参与者、处理和结果。

除了自己做，还有一种方法是与一个有能力的组织合作，例如企业、政府或非政府组织。这样做的优点是这些组织能让你开展那些你自己无法开展的实验。例如，后文有一个实验共有6100万名参与者，这是任何一个研究人员都无法单独实现的规模。合作能让你开展之前不可能开展的实验，但同时也会限制你。例如，大多数公司都不会允许你开展可能有损他们生意或声誉的实验。合作还意味着在你发表研究论文时，可能会被要求“重新调整”研究结果，而且如果你的论文不利于他们的话，有些合作者甚至可能竭力阻止你发表论文。最后，合作还意味着需要成本去发展并维护这些合作关系。

与有能力的组织建立合作关系需解决的核心挑战是找到一个平衡双方利益的方法，巴斯德象限是对此有帮助的一个方法（Stokes 1997）。许多研究人员认为，如果他们研究的是某个组织可能感兴趣的东西，那么他们就不是在做真正的科学研究。这种心态会让成功建立合作关系变得非常困难，而且这也是完全错误的。生物学家路易·巴斯德（Louis Pasteur）的这项开创性的研究可以完美诠释该思维方

式的问题所在。当时巴斯德被一家酒精制造厂邀请去研究将甜菜汁转化为酒精的发酵过程，在此期间，他发现了一种新型微生物，并由此最终提出了疾病细菌学说。新型微生物的发现解决了一个非常实际的问题，既帮助改善了发酵过程，同时又促使科学向前迈了一大步。因此，与其认为有实际用途的研究与真正的科学研究存在冲突，不如将它们看作两个独立的维度。研究的出发点可以是（或不是）实用的，也可以是（或不是）寻求基本的认识。重要的是，有些研究，就比如巴斯德的研究，既是为了解决实际问题也是为了寻求基本认识（图4.17）。属于巴斯德象限中的研究，即本身就具有两个目标的研究，是研究人员和各组织合作的理想之选。在此背景下，我将描述两种合作开展的实验研究：一种是与企业合作，一种是与非政府组织合作。

大型公司，尤其是科技公司，已为开展复杂实验开发了相当精密的、先进的基础设施。在科技行业，这类型实验通常被称为A/B测试，因为它们会比较A、B两项处理的有效性。为了增加广告点击率等目的，这些公司会利用其先进的实验基础设施来开展A/B测试，但这些设施也可以被用来开展促进科学认识的研究。能够说明这类研究可能性的一个例子是脸谱网和加州大学圣迭戈分校的研究人员针对不同信息对选民投票率的影响而合作开展的一项研究（Bond et al. 2012）。

图4.17 巴斯德象限。与其认为研究只能是“基础的”或“应用的”，不如将其看作既是（或不是）为了解决实际问题也是（或不是）为了寻求基本认识。巴斯德针对将甜菜汁转化为酒精的研究，提出了疾病细菌学说，这就是一个既为解决实际问题又为寻求基本认识的研究示例。这类研究是最适合与有能力的组织合作开展的研究。为了解决实际问题而不是寻求基本认识的研究示例是托马斯·爱迪生（Thomas Edison）的研究，为了寻求基本认识而不是解决实际问题的研究示例是尼尔斯·玻尔（Niels Bohr）的研究。改编自Stokes（1997），图3.5。

2010年11月2日美国国会中期选举这天，居住在美国且年龄在18岁及以上的6100万名脸谱网用户参与了一项关于投票的实验。这些用户在访问脸谱网时会被随机分配到三个组中，然后系统会根据分组情况向其信息流顶部推送不同的标语（如果有的话）（图4.18）。

• 对照组；

• 有关投票的信息性消息+一个可点击的“我已投票”按钮+一个计数器（信息组）；

• 有关投票的信息性消息+一个可点击的“我已投票”按钮+一个计数器+已点击“我已投票”的朋友的姓名和照片（信息+社交组）。

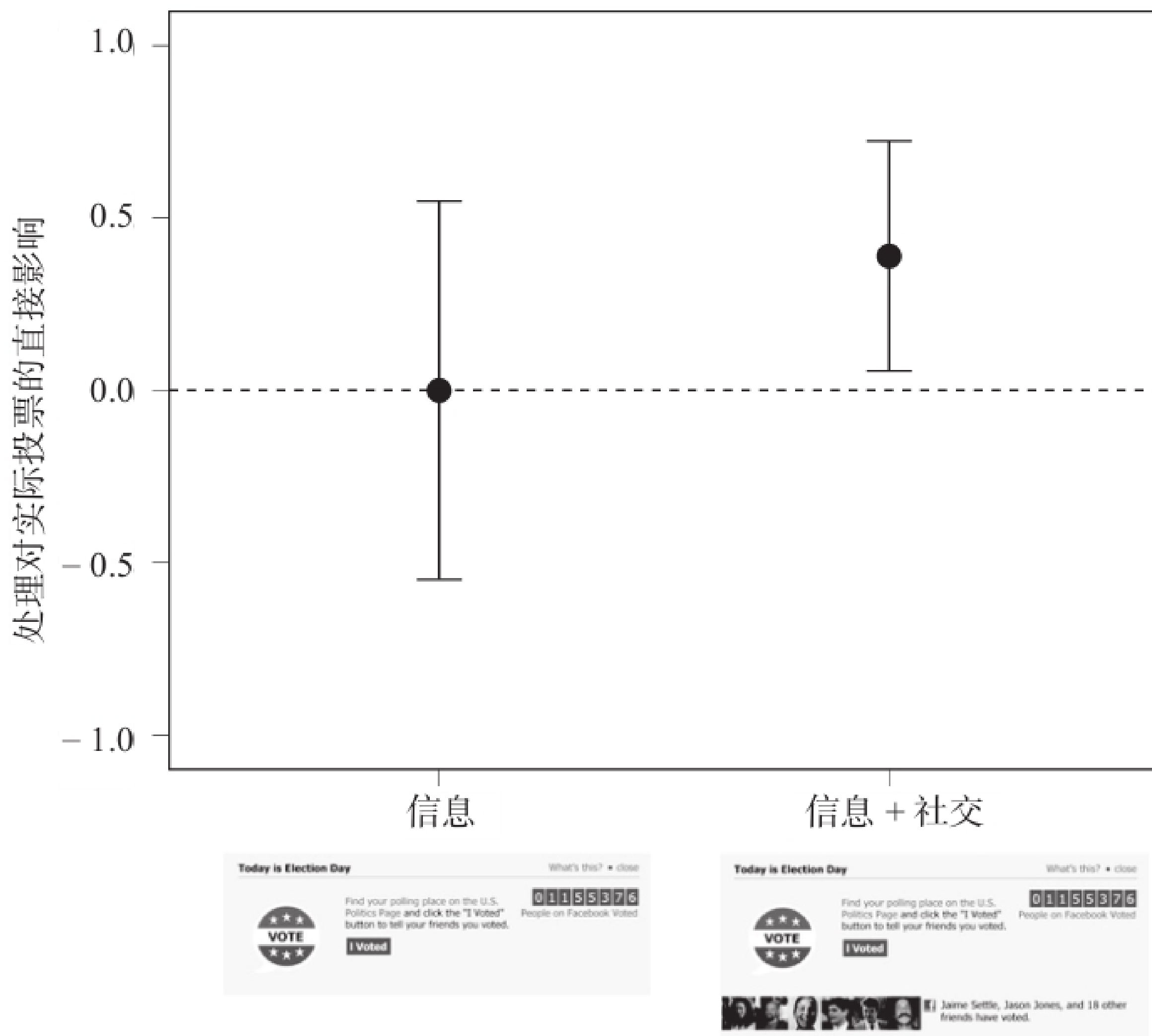


图4.18 脸谱网上一项动员投票的实验结果。信息组的参与者实际投票的概率与对照组的参与者相同，但信息+社交组的参与者实际投票的概率则要稍大一些。竖线代表估算的95%的置信区间。该图中的结果来自与实际投票记录进行匹配的约600万名参与者。改编自Bond et al.（2012），图1。

邦德（Bond）和同事主要研究了两个结果：报告的投票行为和实际的投票行为。首先，他们发现，信息+社交组的人点击“我已投票”的概率比信息组的人高了约2个百分点（约20%∶18%）。其次，研究人员将他们所搜集的约600万名参与者的数据与公开的实际投票记录对比后发现，信息+社交组的人实际投票的概率比对照组的人高0.39个百分点，而信息组的人实际投票的概率则与对照组的人相同（图4.18）。

该实验的这些结果表明，有些在线动员投票的信息要比其他在线动员投票的信息更有效，而且研究人员对该有效性的评估结果还会不同，这取决于他们统计的是报告的投票行为还是实际的投票行为。可惜的是，这项实验并没有提供任何关于社交信息为什么会增加投票率的信息。可能是这些社交信息增加了人们注意到标语的可能性，或者是增加了注意到这一标语的人实际投票的可能性，或者是两个原因都有。因此，这项实验提供了一个有趣的发现，更多的研究人员可能会对其进行探究（参阅例如Bakshy，Eckles，et al. 2012）。

除了推进研究人员的目标以外，这项研究还推进了合作组织（脸谱网）的目标。如果将研究的行为由投票换成购买肥皂，你就会发现，这项研究的结构与衡量在线广告效果的实验结构是完全相同的（可参阅例如Lewis and Rao 2015）。这些针对广告效果的研究经常会衡量接触在线广告对线下行为的影响。因此，这项研究可以提高脸谱网研究在线广告效果的能力，同时有助于脸谱网让潜在的广告商相信脸谱网上的广告在改变行为方面是有效的。

在这项研究中，尽管研究人员和合作组织想要了解的东西基本一致，但也存在一些让他们的合作陷入紧张氛围的分歧，尤其是在参与者的分配上。当时实验中对照组、信息组和信息+社交组的人数分配是非常不均衡的。对研究人员来说，这样不均衡的分配从统计学角度讲很低效，更好的分配方式是每组有1/3的参与者。但之所以当时采取这一分配方式，是因为脸谱网想让每个人都接收信息+社交处理。幸运的

是，研究人员说服脸谱网为信息组和对照组各保留1%的参与者。如果没有对照组，衡量信息+社交的处理效应就基本不可能了，因为这将是一个“干扰观察”实验，而不是随机对照实验。这项研究为与有能力的组织合作开展实验提供了宝贵的实践经验：同样是促成一项实验，有时你需要说服一个组织去实施一项处理，而有时则需要说服一个组织不去实施一项处理（即要求设置一个对照组）。

合作并不总是要找科技公司，也并不总是要开展有数百万参与者的A/B测试。例如，亚历山大·科波克（Alexander Coppock）、安德鲁·格斯（Andrew Guess）和约翰·捷尔诺夫斯基（John Ternovski）就在2016年与非政府环保组织保育选民联盟（League of Conservation Voters）针对不同策略在促进社会动员方面的有效性合作开展了实验。研究人员利用该非政府组织的推特账号发送了公众推文和直接面向个人的、试图侧重于不同类型身份信息的消息，然后测量了这些消息中哪些消息在动员人们签署请愿书和转发有关请愿书的内容方面最为有效。

总的来说，与有能力的组织合作所能达到的实验规模是个人难以达到的，表4.3列出了研究人员与组织合作开展的其他实验实例。此外，合作开展实验要比创建自己的实验简单很多。但合作开展实验也存在缺点：合作会限制参与者、处理及你所能研究的结果的类型，而且还可能引发道德伦理方面的问题。发现合作实验的最好方法是注意到那些能通过有趣的科学研究来解决的实际问题。如果你还不习惯用这样的方式去观察这个世界，那么就很难发现属于巴斯德象限内的实验，但通过练习，你会变得越来越擅长。

表4.3 研究人员和组织合作开展的研究实例

研究主题	参考文献
脸谱网信息流对信息分享的影响	Bakshy, Rosenn, et al.（2012）
部分匿名对在线交友网站上行为的影响	Bapna et al.（2016）
家庭能源报告对用电量的影响	Allcott（2011, 2015）; Allcott and Rogers（2014）; Costa and Kahn（2013）; Ayres, Raseman, and Shih（2013）
应用程序设计对病毒传播的影响	Aral and Walker（2011）
传播机制对传播的影响	Taylor, Bakshy, and Aral（2013）
广告中社交信息的影响	Bakshy, Eckles, et al.（2012）
产品目录更新频率对通过产品目录购买或在线购买的不同客户购买量的影响	Simester et al.（2009）
受欢迎的信息对潜在求职者的影响	Gee（2015）
初始评价对受欢迎程度的影响	Muchnik, Aral, and Taylor（2013）
消息内容对政治动员的影响	Coppock, Guess, and Ternovski（2016）

注意：有些实验的研究人员为合作组织的员工。

4.6 建议

无论你是自行开展实验还是与相关组织合作开展，我都想分享我在工作中发现的特别有用的四条建议，其中前两条适用于任何实验，后两条则主要针对数字时代的实验。

当你要开展一项研究时，我的第一条建议是，在搜集数据前应尽可能多地思考。对习惯于开展实验的研究人员来说，这似乎是显而易见的，但对习惯于利用大数据资源的研究人员来说，这是非常值得注意的（参见第2章）。利用大数据资源时，大多数工作需要在搜集到数据后才能完成，但开展实验是相反的，大多数的工作应该在数据采集前完成。迫使你自己在搜集数据前仔细思考的最好的方法之一是为你的研究创建一个预分析计划，基本描述一下你将进行的分析。

我的第二条适用于所有实验的建议是，没有一个单一的实验是完美的，因此，你应该考虑设计一系列相辅相成的实验。有人将这一策略称为无敌舰队策略。也就是说，与其努力打造一艘庞大的战舰，不如建造许多优势互补的小型战舰。这类多实验研究在心理学领域是很常见的，在其他领域却很少见。幸运的是，有些数字实验的低成本使多实验研究变得更加容易了。

分享完以上两条适用于所有实验的建议后，接下来我将分享两条主要针对数字时代实验设计的建议：创造零可变成本数据和将道德伦理融入设计中。

4.6.1 创造零可变成本数据

开展大型实验的关键是将你的可变成本降低到零。实现这一点最好的方法是自动化和设计有趣的实验。

数字实验可以有截然不同的成本结构，这也使得研究人员能够开展过去不可能开展的实验。考虑这一差异的一种方式是从固定成本和可变成本（实验成本通常被分为这两类）入手。固定成本是指不会随参与者数量的变化而变化的成本。例如，在一项实验室实验中，固定成本可能就是租用场地和购买设备所产生的成本。而可变成本则是指会随参与者数量的变化而变化的成本。例如，在一项实验室实验中，可变成本可能来自给研究助理和参与者的经济报酬。一般来说，模拟实验是固定成本较低，可变成本较高，而数字实验则是固定成本较高，可变成本较低（图4.19）。尽管数字实验的可变成本较低，但如果你愿意尝试将其降低到零，就可以创造出许多令人兴奋的机会。

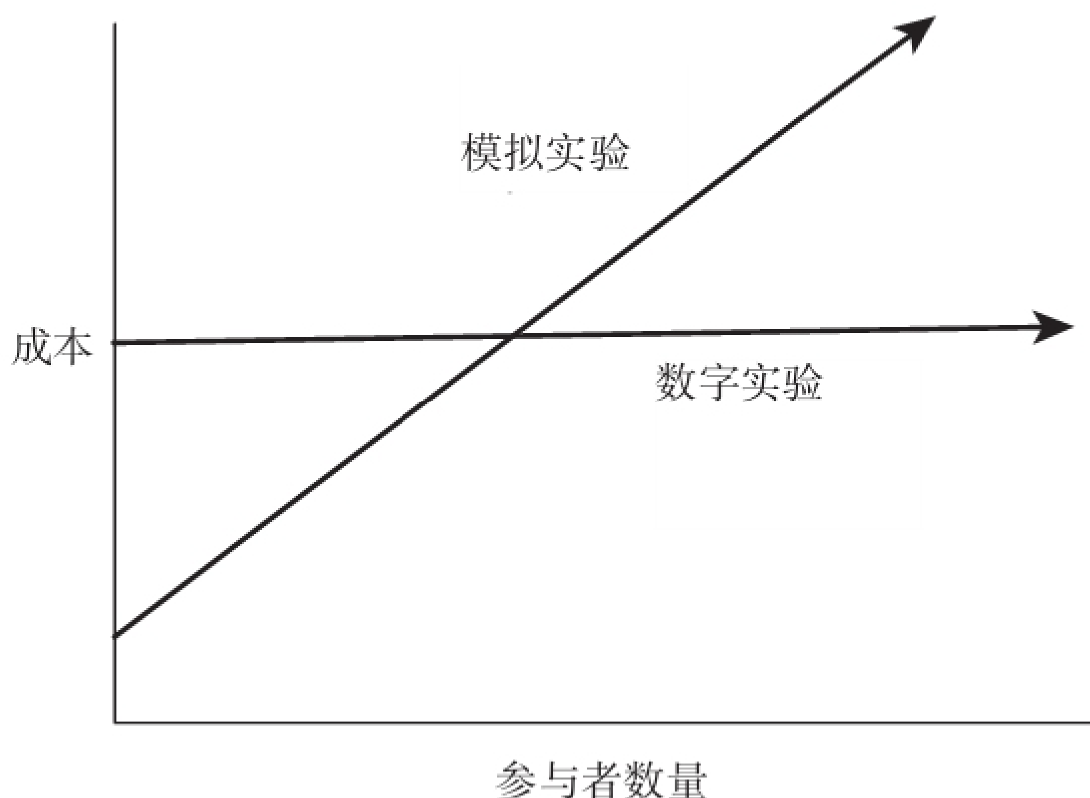


图4.19 模拟实验和数字实验的成本结构示意图。一般来说，模拟实验是固定成本较低，可变成本较高，而数字实验则是固定成本较高，可变成本较低。不同的成本结

构意味着数字实验能达到模拟实验不可能达到的规模。

实验的可变成本主要产生于两个方面，即给研究助理的经济报酬和给参与者的经济报酬，我们可通过不同的策略将这两方面的成本降低到零。给研究助理经济报酬是因为需要他们去招募参与者、实施处理以及测量结果。例如，舒尔茨和同事针对用电量的模拟实地实验就需要研究助理前往每个家庭实施处理并记录电表读数。这就意味着增加参与研究的家庭就会增加成本。而在雷斯蒂沃和范德里杰特针对奖励与人们编辑维基百科贡献的关系而开展的数字实地实验中，增加参与者几乎不会增加任何成本。降低可变行政成本的一般策略是用计算机（便宜的）来代替人工（昂贵的）工作。粗略地说，你可以问自己：当我的研究团队中的每个人都在睡觉时，这个实验还能运转吗？如果答案是肯定的，那么你的实验在自动化方面是很出众的。

实验的可变成本还源于给参与者的经济报酬。一些研究人员会通过使用机器人MTurk和其他在线劳动力市场来减少需支付给参与者的经济报酬。但要想将可变成本降低到零，则需要采取一种不同的方法。很长一段时间以来，研究人员设计的实验都是很枯燥的，所以必须花钱才能让人参与。但如果你能设计一个人们自愿想参与的实验呢？这听起来可能有点令人难以置信，但我自己就曾开展过这样一个实验，我将在后文中对其进行描述，表4.4还列出了其他这类型实验的例子。需要注意的是，设计有趣的实验这一想法与第3章中关于设计更有趣的调查和第5章中关于大规模协作设计的一些主题相呼应。因此，我认为参与者的喜爱度，也可能被称为用户体验，将是数字时代研究设计中越来越重要的一部分。

表4.4 零可变成本的实验实例

用有价值的服务或愉快的体验作为对参与者的酬谢

酬谢方式	参考文献
健康信息网站	Centola（2010）
锻炼计划	Centola（2011）
免费音乐	Salganik, Dodds, and Watts（2006）； Salganik and Watts（2008, 2009b）
有趣的游戏	Kohli et al.（2012）
电影推荐	Harper and Konstan（2015）

如果你想创造零可变成本的实验，就需要确保一切都是全自动化的，而且参与者也不需要任何经济报酬。我将通过我针对文化产品的成功和失败的论文研究来说明如何实现这一点。

我的论文最初是想探究那些令人费解的文化产品的成功。热门歌曲、畅销书和卖座的电影，其人气指数比各自领域的平均水平要高得多。也正因如此，这些产品的市场通常被称为赢家通吃的市场。与此同时，究竟哪首歌、哪本书或哪部电影会成功，却是非常难预测的。编剧威廉·戈德曼（William Goldman）曾对大量学术研究做出了如下的高度概括："当预测成功时，没有人知道任何事。"赢家通吃的市场的不可预测性让我很想知道文化产品的成功多大程度是因为质量，多大程度是因为运气。或者，稍微换种方式表达就是，如果我们可以创造出平行世界并让它们独立运转，那么同样的歌曲在这些世界中也会受欢迎吗？如果不会，造成这些差异的机制又是什么呢？

为了回答这些问题，我和我的论文导师彼得·多兹（Peter Dodds）、邓肯·瓦茨开展了一系列在线实地实验。具体来说，我们创建了一个叫音乐实验室（MusicLab）的网站，人们可以通过该网站发现新的音乐，然后我们利用这一网站进行了一系列实验。我们通过在青少年喜欢的一个网站上发布横幅广告（图4.20）以及媒体广播来招募参与者。进入网站的参与者需先签署知情同意书，然后完成一项简

短的背景问卷，之后会被随机分入对照组或实验组。在对照组中，参与者根据给出的乐队名和歌名自行决定要听哪首歌。他们在听歌的同时会被要求对歌曲进行评分，之后便可以下载这首歌（也可以不下载）。实验组的流程也是一样的，唯一的不同是参与者还可以看到每首歌被前面的参与者下载的次数。此外，实验组的参与者还会被随机分配到8个所谓的平行世界中，每个平行世界都是独立运转的（图4.21）。利用这一设计，我们开展了两个相关的实验。在第一个实验中，我们未对呈现给参与者的表格中的歌曲进行排序，歌曲的受欢迎程度不是很直观。在第二个实验中，我们对呈现给参与者的歌曲进行了排序，歌曲的受欢迎程度更加直观（图4.22）。

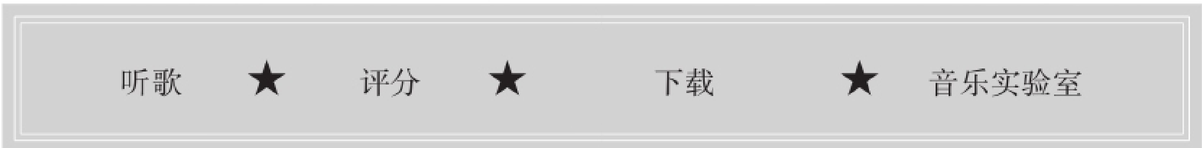


图4.20 我和同事用来为音乐实验室实验招募参与者的横幅广告示例（Salganik, Dodds, and Watts 2006）。经允许复制自Salganik（2007），图2.12。

图4.21 音乐实验室的实验设计。参与者被随机分入对照组或实验组。对照组的参与者需在完全不知道其他参与者的选择的情况下做出选择。而实验组的参与者则被随机分配到8个平行世界，并能看到每首歌在其所在世界中的受欢迎程度（根据前面的参与者的下载次数来衡量），但他们不会看到有关其他世界的任何信息，他们甚至都不知道其他世界的存在。改编自Salganik, Dodds, and Watts（2006），图s1。

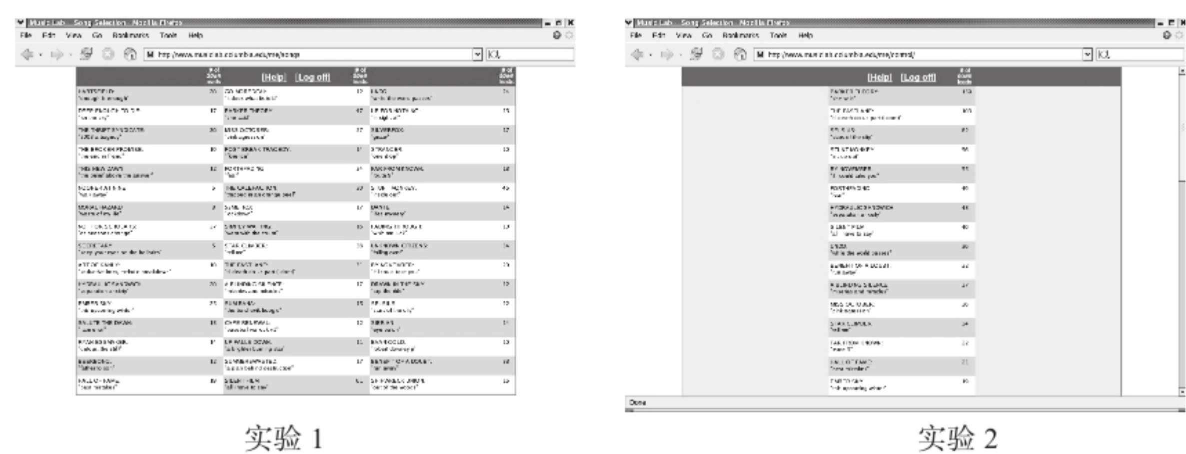


图4.22 音乐实验室实验不同处理的屏幕截图。在实验1的实验组中，研究人员将歌曲及其之前被下载的次数通过一个16×3的矩形表格呈现给参与者，其中每个参与者所看到的歌曲顺序都是随机分配的。在实验2的实验组中，研究人员将歌曲及其之前

被下载的次数按当前受欢迎程度降序排列呈现给参与者。经允许复制自Salganik（2007），图2.7和图2.8。

我们发现每首歌在不同世界的受欢迎程度是不同的，这表明运气在歌曲的成功中起了重要的作用。例如，同样是一个乐队的一首歌，在一个世界中它是48首歌中最受欢迎的，在另一个世界中却排在了第40位。同样的一首歌与相同的另外47首歌竞争，在一个世界中它幸运地成了最流行的歌曲，在其他世界却没有。此外，通过比较两个实验的结果，我们发现实验处理增加了这些市场赢家通吃的性质，这也许说明了技巧的重要性。但纵观所有平行世界后（只有这种设置平行世界的实验才能做到这一点），我们发现实验处理实际上增加了运气的重要性。而且令人惊讶的是，越具吸引力的歌曲，运气对其成功越重要（图4.23）。

音乐实验室实验的设计方式使其基本上能够以零可变成本进行。首先，一切都是全自动化的，所以在我睡觉时它依旧可以进行。其次，给参与者的酬劳是免费的音乐，所以不会产生可变的参与者报酬成本。利用音乐作为酬劳也说明了有时需要在固定成本和可变成本之间进行权衡取舍。利用音乐做实验增加了实验的固定成本，因为我必须花时间从乐队那里获得音乐的使用权，并为他们准备有关参与者对其音乐的反响的报告。但就音乐实验室实验来说，增加固定成本、减少可变成本是正确的做法，因为这使得我们能够开展在规模上约比标准实验室实验大100倍的实验。

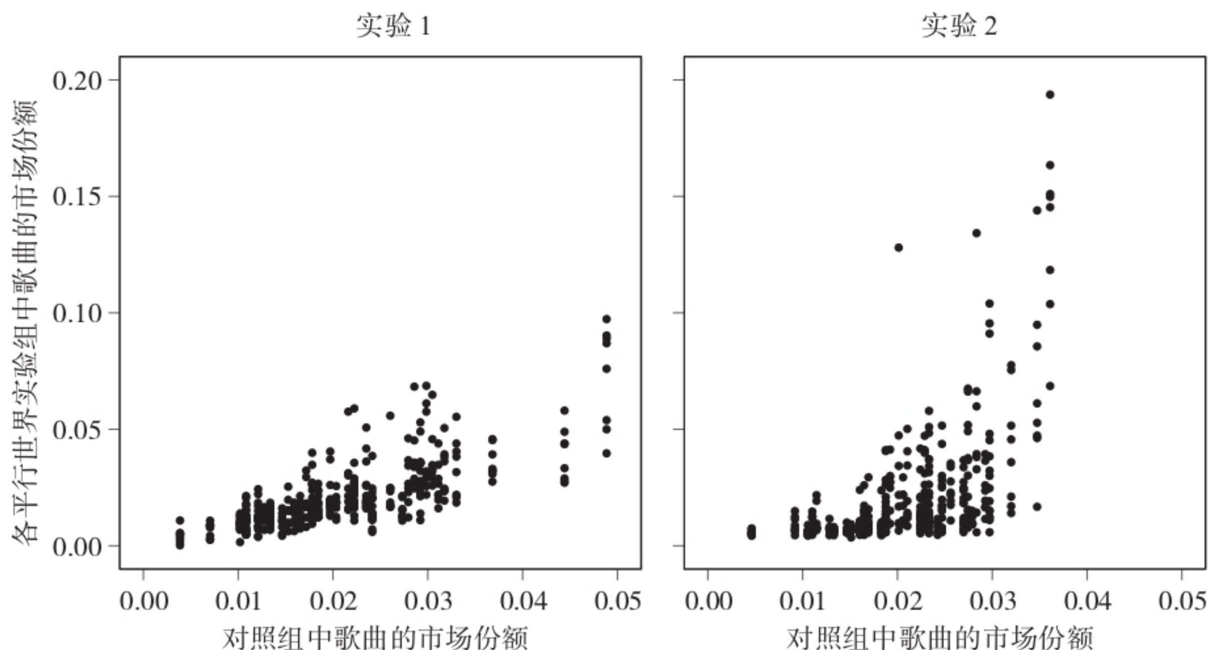


图4.23 表明吸引力和成功之间的关系的音乐实验室实验结果。其中*x*轴是对照组中歌曲的市场份额，作为对歌曲吸引力的一种衡量。*y*轴是8个平行世界的实验组中相同歌曲的市场份额，作为对歌曲成功的一种衡量。我们发现，增强实验处理的力度，特别是将歌曲布局从实验1的方式换成实验2的方式（图4.22），会让成功变得更加不可预测，尤其是对那些最具吸引力的歌曲来说。改编自Salganik, Dodds, and Watts（2006），图3。

此外，音乐实验室实验表明，零可变成本本身并不一定就是最终的目标，它还可以是开展一种新型实验的方式。需要注意的是，尽管我们拥有的参与者数量大约是标准实验室实验参与者数量的100倍，但我们并没有利用所有的参与者来开展100次研究实验处理效应的标准实验室实验。相反，我们做了一些不同的事情，你可以将其看作从心理学实验到社会学实验的转变（Hedström 2006）。我们在实验中关注的是受欢迎程度，这是一个总体结果，而不是个体决定。将关注点转移到总体结果上意味着我们需要大约700名参与者来得出一个单一的数据点（每个平行世界中有700名参与者）。这一实验规模正是因为我们的成本结构才成为可能。总的来说，如果研究人员想要研究个体决定是如何产生总体结果的，那么音乐实验室实验这样的分组实验是非常令

人振奋的选择。在过去，这类实验面临着组织实施上的困难，但因为零可变成本数据成为可能，这些困难也正在逐渐消失。

除了说明零可变成本数据的好处之外，音乐实验室实验还表明了这种方法面临的一个挑战，即高固定成本。就我的事例而言，我非常幸运地遇到了一位名叫彼得·豪塞尔（Peter Hausel）的出色的网站开发人员，然后我们花了大约6个月一起构建了上述实验。也是因为我的导师邓肯·瓦茨获得了一些支持该类研究的资助款项，它才得以实现。我们创建音乐实验室是在2004年，现在的技术已经进步了，所以现在构建这样的实验应该是更加容易了。但高固定成本策略确实是只有那些能够承担得起成本的研究人员才能采用的。

综上所述，数字实验具有与模拟实验截然不同的成本结构。如果你想开展大规模的实验，就应该尝试尽可能降低可变成本，最好能降低到零。你可以通过使实验过程自动化（例如让计算机代替人工）和设计人们自愿想要参与的实验来做到这一点。能够设计出具有这些特征的实验的研究人员，便能开展过去不可能实现的新型实验。但开展零可变成本实验也会引发新的道德伦理问题，这也是我接下来将探讨的主题。

4.6.2 将道德伦理融入你的设计：替代、改进和减少

你可以通过用非实验研究替代实验、改进处理和减少参与者的数量，使实验更加人道。

针对数字实验设计的第二条建议是关于道德伦理的。正如雷斯蒂沃和范德里杰特针对维基百科谷仓之星的实验所示，降低成本意味着道德伦理将成为研究设计中越来越重要的一部分。除了我在第6章将描述的用来指导人体实验的伦理框架之外，设计数字实验的研究人员还

可以参考另一个来源的道德伦理理念：指导动物实验的伦理原则。尤其是罗素（Russell）和伯奇（Burch）在其具有里程碑意义的《人道实验技术原则》（*Principles of Humane Experimental Technique*）一书中提出的指导动物实验的三个原则：替代、改进和减少。我认为，这三个原则稍做修改后也可以被用来指导人体实验的设计。

- 替代：如果可能的话，用侵害性更小的方法来替代实验。
- 改进：改进处理，使其尽可能不具危害性。
- 减少：尽可能减少实验的参与者。

我首先将通过一个引发伦理争议的在线实地实验来更具体地阐明这三个原则，并向大家展示它们是如何造就更好、更人道的实验设计的。然后，我将描述这三个原则如何让实验设计人员想到具体实用的优化方案。

最具伦理争议的数字实地实验之一是亚当·克雷默（Adam Kramer）、杰米·吉约里（Jamie Guillory）和杰弗里·汉考克（Jeffrey Hancock）开展的一项被称为“情绪感染”的实验。实验是在脸谱网上进行的，出发点是为了解决一些科学和实际问题。当时，脸谱网用户与脸谱网互动的主要方式是信息流，即通过算法控制的来自脸谱网用户好友的状态更新。一些批评脸谱网的人认为，因为信息流中大多是积极的帖子，比如发一下最近的聚会，所以可能会让用户觉得自己的生活似乎不如朋友的精彩，进而感到难过。但也可能恰好相反，也许看到你的朋友玩得开心会让你也感到快乐。为了验证这两个相互矛盾的假设并进一步了解一个人的情绪如何被其朋友的情绪所影响，克雷默和同事开展了一项实验。在这项为期一周的实验中，他们将大约70万名用户分成了4组：一个是“消极减少”组，研究人员会随机屏蔽含有消极词汇（例如“难过”）的帖子，以免其出现在这些人的信息流中；一个是“积极减少”组，研究人员会随机屏蔽含有积

极词汇（例如“开心”）的帖子，以免其出现在这些人的信息流中；另外两组是对照组。在“消极减少”组的对照组中，研究人员会以与“消极减少”组相同的屏蔽率来随机屏蔽帖子，但不会考虑帖子是消极的还是积极的。在“积极减少”组的对照组中，研究人员进行了相同的操作。该实验设计表明，适当的对照组并不总是一个不对其进行任何处理的组。也就是说，有时为了进行精确的比较以解答所研究的问题，研究人员需对对照组也实施一定的处理。4个组的用户可以通过信息流以外的其他脸谱网功能看到被屏蔽的其好友的帖子。

克雷默和同事发现，就“积极减少”组的参与者来说，他们更新状态中积极词汇的比例下降了，消极词汇的比例则上升了。而就“消极减少”组的参与者来说，他们更新状态中积极词汇的比例上升了，消极词汇的比例则下降了（图4.24）。但这些处理的效应量是很小的：实验组和对照组在积极词汇和消极词汇上出现差异的概率是千分之一。

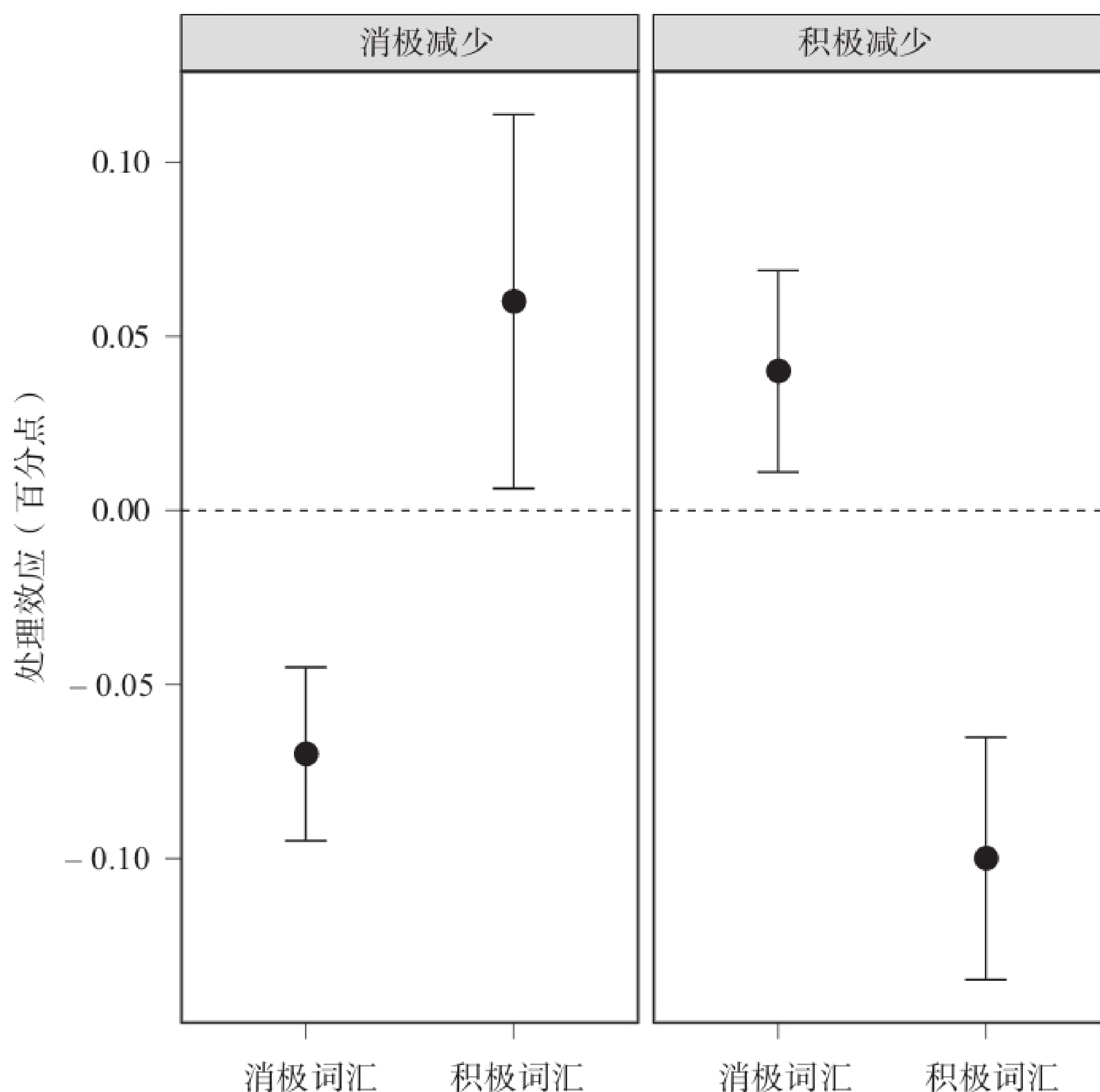


图4.24 情绪感染的证据。“消极减少”组的参与者使用的消极词汇较少，积极词汇较多；“积极减少”组的参与者使用的消极词汇较多，积极词汇较少。竖线代表估算的标准误差。改编自Kramer，Guillory，and Hancock（2014），图1。

在探讨该实验引发的道德伦理问题之前，我想用本章前面的一些概念来描述三个科学问题。

首先，我们还不清楚该实验的实际细节与相关理论是如何关联的。换句话说，我们对该实验的构念效度尚有疑问。我们还不清楚积极词汇和消极词汇的数量是否能很好地反映参与者的情绪状态，因为

我们不清楚：（1）人们在帖子中使用的词汇是否能很好地反映他们的情绪，（2）研究人员采用的特定的情绪分析技术是否能准确地推断情绪（Beasley and Mason 2015; Panger 2016）。换句话说，词汇可能并不是一个很好的情绪指标，上述特定的情绪分析技术也可能并不是一个很精确的测量方法。

其次，该实验的设计和分析并没有告诉我们谁是受影响最大的（即没有对处理效应的异质性进行分析），也没有告诉我们可能的原理。在该事例中，研究人员有很多关于参与者的信息，但在分析实验时基本上没有考虑这些信息。

最后，该实验中的效应量是很小的：实验组和对照组出现差异的概率大约是千分之一。在他们的论文中，克雷默和同事表示，这样的效应量也是很重要的，因为每天会有数以亿计的人查看他们的信息流。换句话说，他们认为，即使对每个人的影响很小，但对数以亿计的人的影响总和很大。就算你同意他们这个观点，这样的效应量对情绪传播这样一个更为普遍的科学问题是否具有重要意义，我们仍然无从得知（Prentice and Miller 1992）。

除了这些科学问题，克雷默和同事的论文在《美国国家科学院院刊》（*Proceedings of the National Academy of Sciences*）上发表后没几天就引来了研究人员和媒体的强烈抗议（具体争论的观点我将在第6章更详细地描述）。这场争论中提出的问题致使上述院刊罕见地就有关这项实验的道德伦理问题和伦理审查过程的担忧发表了社论（Verma 2014）。

描述完情绪感染这一实验，现在我想说明的是，上述三个原则能为真正的研究带来具体实用的改进方案（无论你个人如何看待这个特定实验的道德伦理问题）。第一个原则是替代：如果可能的话，研究人员应该设法用侵害性、风险性更小的方法来替代实验。例如，研究人员可以尝试利用自然实验，而不是开展随机对照实验。正如第2章所

描述的，自然实验是指现实世界中发生的事情刚好大致满足了对处理的随机分配（例如，抽签决定入伍人选）。自然实验的伦理优势在于，研究人员不必实施处理，因为环境会替他们实施。例如，几乎就在克雷默等人开展情绪感染实验的同一时间，科维略（Coviello）和同事发现，人们在下雨天发的帖子中消极词汇会比较多、积极词汇会比较少，因此，通过利用天气的随机变化，他们可以在不进行干预的情况下研究信息流变化的影响，这就好像是天气在替他们开展实验一样，他们的这一实验可被称为情绪感染自然实验。他们的具体步骤有点复杂，但我们现在探讨的是如何用侵害性、风险性更小的方法来替代实验，所以他们的事例告诉我们最重要的一点是，通过利用自然实验，科维略和同事可以在不开展自己的实验的情况下了解情绪传播。

第二个原则是改进：研究人员应该设法改进实验处理，以使其尽可能无害。例如，研究人员可以增加积极或消极的内容，而不是屏蔽积极或消极的内容。增加内容的设计会改变参与者信息流的情绪内容，但也能解决批评者提出的一个顾虑，即实验可能导致参与者错过信息流中的重要信息。在克雷默和同事的设计中，重要信息与不重要信息被屏蔽的概率是一样的。但在增加内容的设计中，被取代的将会是那些不太重要的信息。

第三个原则是减少：研究人员应该设法将实验参与者的数量减少到完成科学目标所需要的最小数目。在模拟实验中，因为参与者的可变成本较高，所以研究人员自然会这么做。但在数字实验中，尤其是零可变成本的实验中，研究人员不会因为实验的规模而面临成本问题，这就有可能导致不必要的大规模实验。

例如，克雷默和同事就应该利用有关参与者的处理前信息，比如处理前的发帖行为，来使他们的分析更加高效。更具体地说，克雷默和同事应该比较实验组中积极词汇的比例变化和对照组中积极词汇的比例变化，而不是比较实验组的积极词汇比例和对照组的积极词汇比

例。前者有时被称为混合设计（图4.5），有时也被称为双重差分估计量法。也就是说，研究人员应该计算出每个参与者的变化量（处理后行为—处理前行为），然后比较实验组和对照组参与者的变化量。这种双重差分的方法从统计学角度来讲会更加高效，因为它意味着研究人员可以利用更小的样本量实现相同的置信度。

就克雷默和同事的实验而言，因为没有原始数据，所以我们很难确切地知道双重差分估计量法的效率比原来方法的效率具体高出了多少，但可以通过其他相关的实验获得一个大致的概念。邓（Deng）等人报告称，通过采用其中一种双重差分估计量法，他们将三个不同在线实验的估计量的方差减少了约50%。谢（Xie）和奥里塞（Aurisset）也报告了类似的结果。50%的方差缩减意味着如果克雷默和同事采用一种稍微不同的分析方法，那么他们所需的样本量可能就能减少一半。换句话说，稍微改变一下分析方法或许就能减少35万名参与者了。

在这一点上，你可能会想研究人员为什么要在意这35万人在情绪感染实验中是否多余呢？这是因为情绪感染实验的两个特性使我们有必要担忧实验规模过大，许多数字实地实验都具有这两个特征：（1）不确定实验是否会对至少某些参与者产生伤害，（2）参与者不是自愿参与实验的。尽可能缩小具有这些特征的实验的规模似乎是合理的。

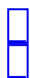
要澄清的一点是，上述减小实验规模的要求并不意味着你不应该开展大规模零可变成本的实验。这只是意味着，实验规模刚好能够让你实现科学目标就可以了。确保实验规模合适的一个重要方法是进行功效分析（Cohen 1988）。在模拟时代，研究人员通常会通过功效分析来确保他们的研究规模不会太小（即参与者过少）。但现在，研究人员应该通过功效分析来确保他们的研究规模不会太大（即参与者过多）。

综上所述，替代、改进和减少这三个原则有助于研究人员将道德伦理融入实验设计中。当然，上述每种改善情绪感染实验的方案，其本身也是瑕瑜互见。例如，通过自然实验得来的证据并不总是像通过随机实验得来的那样纯粹，而且增加内容从逻辑上讲可能比屏蔽内容更难以实现。因此，提出这些改善方案并不是为了事后批评其他研究人员的决定，而是为了说明如何在现实情况下应用这三个原则。事实上，在研究设计中，权衡取舍的问题一直都存在，而且在数字时代，这些权衡将越来越涉及道德伦理方面的考虑。稍后在第6章，我将给出一些有助于研究人员理解和探讨这些权衡取舍的原则以及道德伦理框架。

4.7 结论

数字时代使研究人员能够开展以前不可能的实验。研究人员不仅可以开展大规模的实验，还可以利用数字实验的特殊性质提高效度，评估处理效应的异质性，以及弄清楚原理。这些实验可以在完全数字化的环境中进行，也可以在物理世界中使用数字设备来进行。

正如本章所示，这些实验可以与有能力的公司合作开展，也可完全由研究人员自行开展。而且并不是只有在大型科技公司工作的人才能开展数字实验。如果要自己设计实验，你可以尝试将你的可变成本降低到零，并通过替代、改进和减少这三个原则，将道德伦理融入设计中。研究人员对数百万人生活的干预能力在不断增强，这就意味着我们同样应该越来越关注研究设计是否符合道德伦理要求。能力越大，责任越大。



第5章　进行大规模协作

5.1 简介

维基百科是不可思议的。志愿者的大规模协作创造了这样一部每个人都能使用的百科全书。维基百科成功的关键不是新知识，而是新的协作形式。幸运的是，数字时代使许多新形式的协作成为可能。因此，我们现在应该想想：有哪些重大的科学问题，我们单独无法解决的问题，是现在通过协作能够解决的？

当然，科研协作已经不是什么新鲜事了，新鲜的是数字时代使我们能够与数量更多、更多样化的人进行协作：全世界能够上网的数十亿人。我预计这些新的大规模协作将产生惊人的结果，不仅是因为它们所涉及的人数，还因为这些人所具有的多种多样的技能和视角。我们如何才能让每个互联网用户都参与我们的研究过程呢？如果你有100名研究助理，你能做什么？如果有10万名熟练的协作者呢？

现在有许多种形式的大规模协作，计算机科学家通常会根据技术特点把它们分成大量的类别（Quinn and Bederson 2011）。然而在本章中，我将根据它们能如何被用于社会研究来分类。具体来说，我认为将它们大致分为三种类型的项目是有帮助的：人本计算、公开募集和分布式数据采集（图5.1）。

图5.1 大规模协作示意图。本章将围绕三种主要形式的大规模协作展开论述：人本计算、公开募集和分布式数据采集。更一般地讲，大规模协作将源于公众科学、众包和集体智慧等领域的想法结合了起来。

我将在本章后面部分更详细地描述每一类型的项目，但现在先让我简单描述一下每个类型。人本计算项目非常适合任务简单、数量庞大的问题，例如给100万张图片贴上标签。这类项目在过去可能是由作

为研究助理的大学生来完成的。为这类项目做出贡献不需要具备与任务相关的技能，并且最终的输出通常是所有贡献的平均值。人本计算项目的一个经典示例是星系动物园（Galaxy Zoo），在这个项目中，10万名志愿者协助天文学家对100万个星系进行了分类。公开募集项目则非常适合为明确公式化的问题寻找新奇的、令人意想不到的答案。这类问题在过去可能需要请教同行。为这类项目做出贡献需要具备与任务相关的专业技能，且最终的输出通常是所有贡献中最好的那一个方案。公开募集项目的一个经典示例是网飞奖，在这个项目中，成千上万的科学家和黑客致力于开发新的算法来预测用户对电影的评价。最后，分布式数据采集项目非常适合大规模的数据采集。这类项目在过去可能是由作为研究助理的大学生或调查研究公司来完成的。为这类项目做出贡献通常能够接触到研究人员无法接触到的数据采集点，且最终的产品就是所有贡献的简单合并。分布式数据采集项目的一个经典示例是观鸟数据库（eBird），在这个项目中，成千上万的志愿者会贡献有关他们所看到的鸟类的报告。

大规模协作在天文学（Marshall, Lintott, and Fletcher 2015）和生态学（Dickinson, Zuckerberg, and Bonter 2010）等领域有着丰富而悠久的历史，但在社会研究中还不常见。但是，通过描述其他领域的成功项目并提供一些关键的组织原则，我希望能让大家相信两件事。首先，大规模协作可以被用于社会研究。其次，使用大规模协作的研究人员将能够解决以前似乎不可能解决的问题。虽然倡导大规模协作的人经常会说采用大规模协作能够省钱，但其实它的优点远不止于此。正如我将要展示的，大规模协作不仅能让我们以更便宜的方式开展研究，还能让我们把研究做得更好。

在本章之前的章节中，大家已经明白了用下述三种不同的方式与人互动能了解到什么：观察他们的行为（第2章）、向他们提问（第3章）和招募他们来参与实验（第4章）。在本章中，我将向大家展示，做我们的研究协作者能了解到什么。对于上述三种主要形式的大规模

协作，每一种我都将描述一个典型示例，然后再进一步用其他示例来阐明与之相关的重要知识，最后描述这种大规模协作被用于社会研究的可能方式。本章最后将介绍5个有助于你设计自己的大规模协作项目的原则。

5.2 人本计算

人本计算要解决的是很宏大的问题。我们需要先将每个问题分解成多个简单的问题，然后将分解后的问题发布给志愿者，最后再将结果整合。

在人本计算项目中，每个人接到的都是简单的小任务，但将所有人的成果整合后就能解决那些对一个人来说过于宏大的问题。如果曾有一个研究问题让你产生了这样的想法——如果我有1000名研究助理，就能解决这个问题，那么这个问题就适合用人本计算来解决。

人本计算项目的一个典型示例是星系动物园。在这个项目中，超过10万名志愿者对大约100万个星系的图像进行了分类，而且其准确度与早期对星系图像进行分类的专业天文学家的准确度差不多，但天文学家所分类的图像的数量要少得多。这次大规模协作完成了对更多图像的分类，进而使天文学家对星系是如何形成的有了新的发现，他们还发现了一个被称为“绿豌豆”（Green Peas）的全新星系。

尽管星系动物园似乎与社会研究不太沾边，但其实在很多情形下，社会研究人员也想对图像或文本进行编码、分类或标示。在有些情况下，这类分析可以利用计算机来完成，但有些形式的分析对计算机来说很难，而对人来说很简单。正是这些对人来说简单、对计算机来说很难的微任务，可以通过人本计算来完成。

不仅是星系动物园中的微任务很具普遍性，该项目的结构也很具普遍性。星系动物园以及其他人本计算项目通常都采用分解－运用－整合这一策略（Wickham 2011），一旦你理解了这个策略，就能用它

来解决许多问题。首先，一个大问题会被分解成许多小问题。然后，运用人力来独立解决每个小问题。最后，将人力解决的结果整合，形成一个一致的解决方案。在此背景下，让我们来看看分解–运用–整合这一策略在星系动物园中是如何被使用的。

5.2.1 星系动物园

星系动物园凝聚众多志愿者之力对100万个星系进行了分类。

星系动物园是2007年牛津大学天文学研究生凯文·肖文斯基（Kevin Schawinski）为解决他所面临的一个问题而想到的。简单来说就是，肖文斯基对星系很感兴趣，而人们可以根据星系的形态（椭圆形或螺旋形）和颜色（蓝色或红色）对它们进行分类。当时，天文学家的传统观点是，像我们的银河系这样的螺旋星系的颜色是蓝色的（意味着年轻），而椭圆星系的颜色则是红色的（意味着年老）。肖文斯基对这一传统观点有所怀疑。他猜想，尽管上述规律一般来讲是正确的，但也可能有相当数量的星系例外，通过研究这些不符合预期规律的不同寻常的星系，他便能对星系的形成过程有所了解。

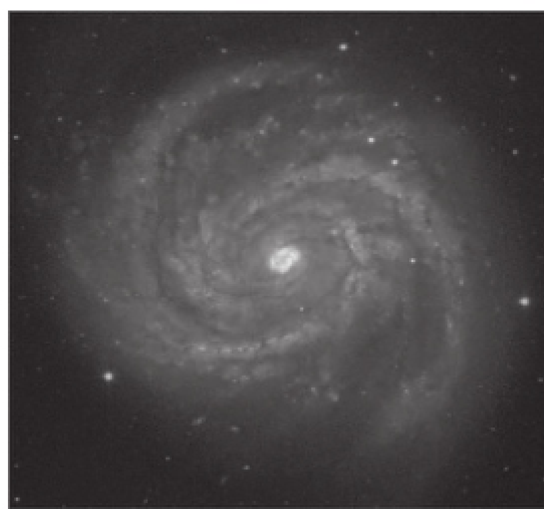
因此，为了推翻传统观点，肖文斯基需要的是大量按形态分类的星系，也就是已被划分为螺旋形或椭圆形的星系。但问题是现有的分类算法还不足以用于科学研究。换句话说，在当时，分类星系对计算机来说是一项很难的任务。因此，肖文斯基需要的是大量人工分类的星系。带着研究生的热忱，肖文斯基开始了分类工作。经过7天每天12小时的马拉松式奋战，他共对5万个星系进行了分类。尽管5万个星系听起来可能很多，但其实仅占斯隆数字天空勘测计划（Sloan Digital Sky Survey）所拍摄的将近100万个星系的大约5%。肖文斯基意识到他需要采取一个更具扩展性的方法。

幸运的是，对星系进行分类并不需要人们接受有关天文学的高深培训，你可以很快就教会一个人如何进行星系分类。换句话说，尽管星系分类对计算机来说是一项困难的任务，对人来说却是非常简单的。因此，当肖文斯基和同学克里斯·林托特（Chris Lintott）坐在牛津大学附近的一家小酒馆时，他们想到了创建一个网站，然后让志愿者对星系的图像进行分类。几个月后，星系动物园诞生了。

在星系动物园的网站上，志愿者需接受几分钟的训练。例如，了解螺旋星系和椭圆星系的区别（图5.2）。完成该训练后，每个志愿者必须通过一项相对简单的考试，即正确地将15个星系中的11个进行分类，然后就能开始通过一个简单的基于网络的界面（图5.3）对未知星系进行真正的分类了。从志愿者到天文学家的转变在不到10分钟内就完成了，而且只需跨越最低的门槛——一个简单的测试。



椭圆星系



螺旋星系

图5.2 椭圆星系和螺旋星系这两类主要星系的示例。星系动物园通过10多万名志愿者对约100万张图像进行了分类。经允许复制自GalaxyZoo.org和斯隆数字天空勘测计划。

在一份报纸报道了该项目之后，星系动物园吸引来了第一批志愿者，然后在大约6个月的时间里，项目吸引了超过10万名公民科学家，人们之所以参与是因为他们喜欢这项任务，并且想要帮助推进天文学

的发展。这10万名志愿者总共贡献了4000多万条分类信息，其中大部分分类信息是由一些相对较少的核心参与者完成的（Lintott et al. 2008）。

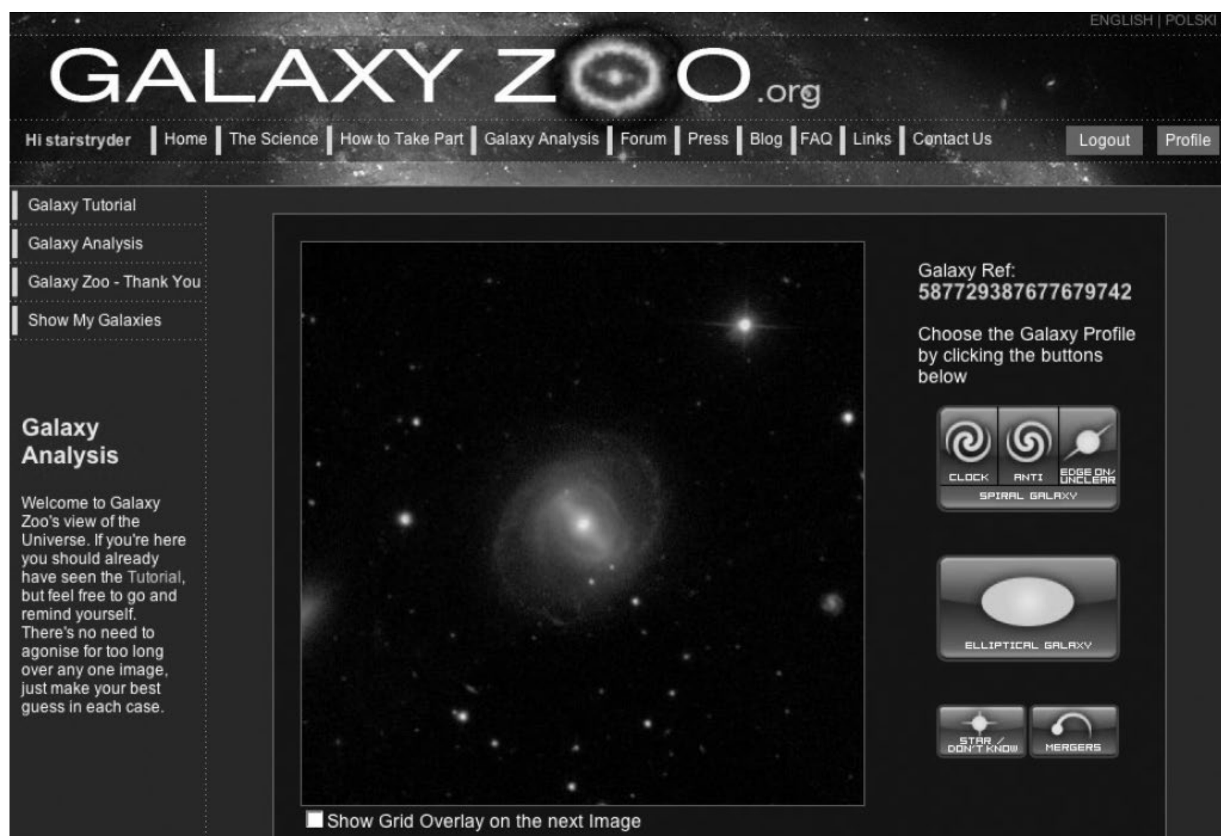


图5.3 志愿者被要求对单个图像进行分类的输入屏。根据斯隆数字天空勘测计划所拍摄的一张图片，经克里斯·林托特允许复制而来。

有雇用大学生做研究助理经验的研究人员可能会立刻对数据质量产生怀疑。尽管这种怀疑是合理的，但星系动物园这一项目表明，志愿者所贡献的数据经正确地清洗、消除偏差和整合后也能产生高质量的结果（Lintott et al. 2008）。让公众创造出专业质量数据的一个重要技巧就是重复开展，即让许多不同的人来执行相同的任务。在星系动物园这一项目中，志愿者对每个星系都贡献了大约40条分类数据，这种程度的重复率是雇用大学生做研究助理的研究人员永远都无法达到的，因为他们需要更加关注每个个体分类数据的质量。志愿者用重复弥补了他们在训练方面的欠缺。

尽管不同志愿者对同一星系进行了多重分类，但要想把这些多重分类数据结合起来以得出一个一致的分类还是很困难的。鉴于大多数人本计算项目遇到的挑战都非常相似，所以简要回顾一下星系动物园研究人员得出一致分类的三个步骤很有帮助。

首先，研究人员通过删除虚假分类清洗了数据。例如，那些反复对同一星系进行分类（如果有人试图操纵结果可能就会这么做）的人会被删除所有的分类数据。这类清洗及其他类似清洗所删除的数据约占总分类数据的4%。

然后，研究人员需要消除清洗后的分类数据中的系统偏差。通过一系列嵌入在原始项目中的偏差检测研究，例如给一些志愿者呈现黑白的而不是彩色的星系图像，研究人员发现了多个系统偏差，例如把因距离遥远而外表模糊的螺旋星系划分成椭圆星系（Bamford et al. 2009）。调整这些系统偏差非常重要，因为重复分类也无法自动消除系统偏差，重复分类只是有助于消除随机误差。

最后，研究人员需要用一种方法把消除偏差后的个体分类数据结合起来，得出一个一致的分类。将每个星系的所有分类数据结合起来的最简单的方法是选择最常见的分类数据。但这就意味着每个志愿者的权重是一样的，而研究人员则认为有些志愿者要比其他志愿者更擅长分类。因此，他们开发了一个更复杂的迭代加权程序，试图检测出最好的分类数据并赋予它们更多的权重。

因此，经过清洗、偏差消除和加权这三个步骤后，星系动物园的研究团队将4000万条志愿者的分类数据转化成了一组一致的形态分类。在将这些形态分类与之前三次由专业天文学家进行的小规模星系分类（其中包括为星系动物园的诞生贡献了部分灵感的肖文斯基所进行的分类）进行比较后发现，它们之间的吻合度很高。因此，将志愿者的贡献整合后也能得出高质量的分类数据，而且其规模是研究人员个人无法企及的（Lintott et al. 2008）。事实上，通过对如此大量

的星系进行人工分类，肖文斯基、林托特和其他相关人员发现，大约只有80%的星系是符合预期规律的，即螺旋星系是蓝色的、椭圆星系是红色的，许多论文都提及了这一发现（Fortson et al. 2011）。

至此，大家应该明白了星系动物园是如何遵循分解－运用－整合这一策略的，这一策略也被用于大多数人本计算项目。首先，把一个大问题分解成多个小问题。以星系动物园为例就是，对100万个星系进行分类这个大问题被分解成了100万个对一个星系进行分类的小问题。然后，分别运用人力对每个小问题进行操作。以星系动物园为例就是，志愿者将每个星系划分为螺旋形或椭圆形。最后，整合所有结果以得出一个一致的结果。以星系动物园为例就是，通过清洗、偏差消除和加权，得出每个星系的一致分类。尽管大多数项目都会采用这一通用的策略，但需要根据所处理的具体问题对每个步骤进行相应的调整。例如，下面这个人本计算项目采用的也是这个策略，但运用和整合这两个步骤是截然不同的。

对星系动物园的团队来说，这个项目仅仅是个开始。很快他们便意识到，尽管能对将近100万个星系进行分类，但这个规模还不足以配合新的大约能拍到100亿个星系的数字天空勘测（Kuminski et al. 2014）。要想应对从100万到100亿的增长，也就是10000倍的增长，他们需要招募的参与者数量大致是星系动物园这个项目的10000倍。尽管互联网上的志愿者很多，但也不是无限的。因此，研究人员意识到，如果他们想要处理数据量日益增多的数据，就需要一个新的、更具扩展性的方法。

于是，曼达·班纳吉（Manda Banerji）同肖文斯基、林托特和星系动物园团队的其他成员一起，开始教计算机进行星系分类。更具体地说就是，班纳吉利用星系动物园所创建的人工分类数据建立了一个机器学习模型，它能够根据图像特征预测星系的人工分类结果。如果

该模型的预测精度很高的话，那么星系动物园的研究人员基本就能用它对无限的星系进行分类了。

班纳吉和同事方法的核心实际上和社会研究中常用的技术非常相似，尽管乍一看这种相似性可能并不明显。首先，班纳吉和同事将每张图像转化成了一组概括其性质的数字特征。例如，星系图像可以有三个特征：图像中蓝色像素点数量、像素亮度方差、非白像素比例。选择正确的特征是这一问题的重要部分，这通常需要专业领域的专业知识。这第一步通常被称为特征工程，经过这一步，班纳吉和同事构建了一个数据矩阵，每一张图集都由一行和三列数据描述。根据该数据矩阵和期望输出值（例如，如采用人工分类，某一图像是否会被划分为椭圆星系），研究人员可以构建一个统计或机器学习模型，例如逻辑回归，以此根据图像的特征预测人工分类的结果。最后，研究人员可以利用该统计模型中的参数预测出新的星系的分类（图5.4）。在机器学习中，这种利用标签示例创建一个能标记新数据的模型的方法被称为监督式学习。

班纳吉和同事的机器学习模型的特征，比我下面这个虚构的小例子的特征要复杂得多。在这个例子中，研究人员选择“德伏古勒轴比”（de Vaucouleurs fit axial ratio）这样的性质作为特征，使用的模型也不是逻辑回归，而是一个人工神经网络。利用选择好的特征、模型和星系动物园的一致分类，她算出了每个特征的权重，然后利用这些权重对星系的分类进行预测。例如，她经过分析发现，“德伏古勒轴比”较低的图像更有可能属于螺旋星系。有了这些权重，她便能相对准确地预测一个星系的人工分类结果了。

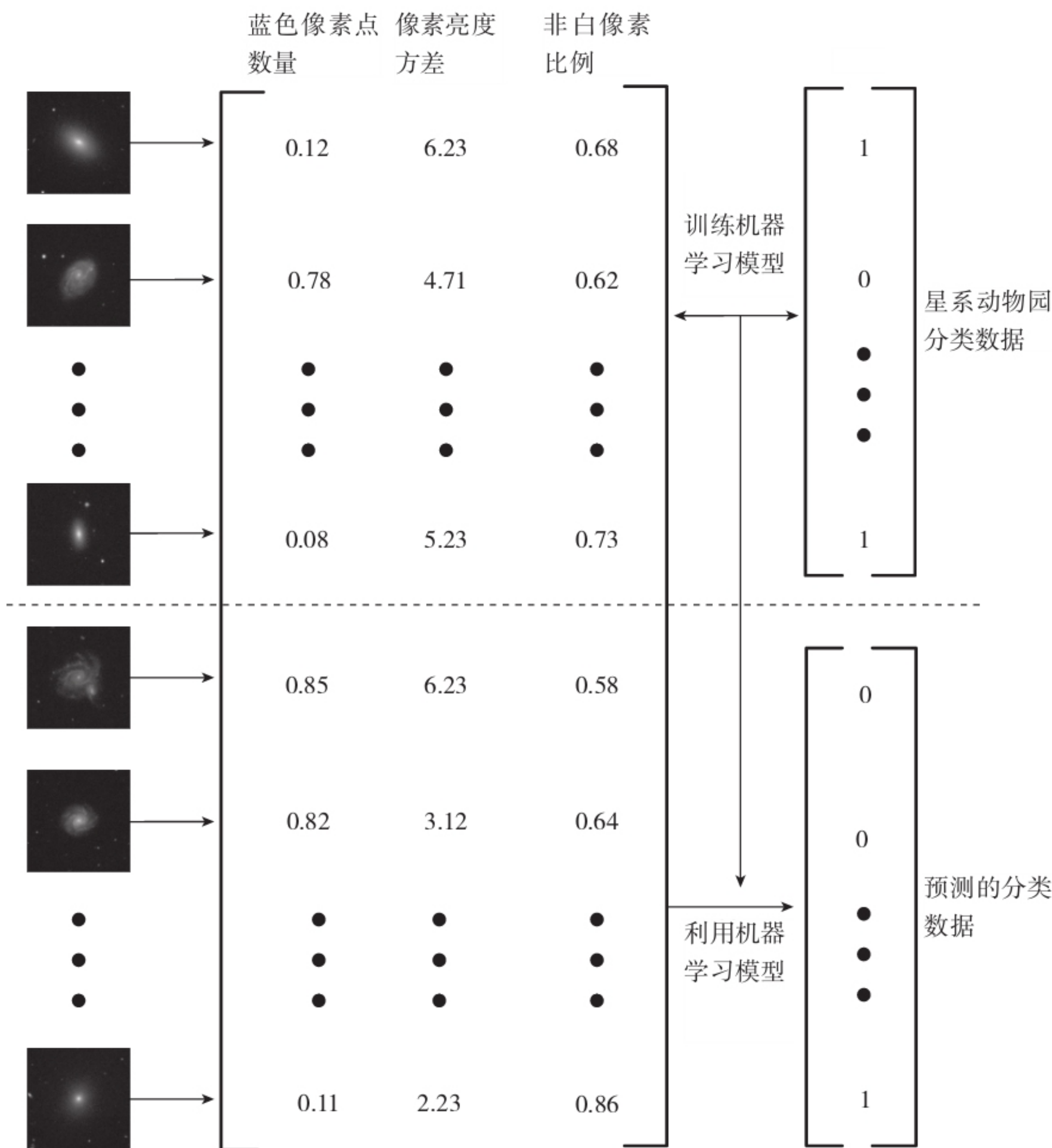


图5.4 班纳吉等人利用星系动物园的分类数据训练一个机器学习模型进行星系分类的简化示意图。星系图像被转换成了一个特征矩阵。在这个简化的例子中，星系图像有三个特征（图像中蓝色像素点数量、像素亮度方差、非白像素比例）。然后，他们利用星系动物园的分类数据训练一个机器学习模型。最后，他们利用这个机器学习模型预测星系动物园分类数据中未涉及的其他星系的分类。我称之为计算机辅助人本计算项目，因为它不是让人来解决问题，而是让人构建一个数据集，然后用这个数据集训练计算机来解决问题。这种计算机辅助人本计算系统的优点是，它能让你用有限的人力处理几乎无限的数据。星系图像经允许复制自斯隆数字天空勘测计划。

班纳吉和同事的工作让星系动物园变成了我所说的计算机辅助人本计算系统。对这类混合系统最好的解读方式是，它们不是让人来解决问题，而是让人构建一个数据集，然后用这个数据集训练计算机来解决问题。有时，训练计算机来解决问题需要大量的示例，而获得足够数量示例的唯一方法就是大规模协作。这种计算机辅助方法的优点是它能让你用有限的人力处理几乎无限的数据。例如，一个拥有100万个星系的人工分类数据的研究人员可以构建一个预测模型，然后用这个模型对10亿甚至10000亿个星系进行分类。如果星系的数量很庞大，那么这种人与计算机的混合系统将是唯一可能的解决办法。然而这种无限的可扩展性并不是没有成本的。构建一个能够正确预测人工分类结果的机器学习模型本身就是一个难题，但幸好已经有一些针对这一主题的优秀书籍了（Hastie, Tibshirani, and Friedman 2009; Murphy 2012; James et al. 2013）。

星系动物园很好地表明了许多人本计算项目是如何逐步发展的。首先，一个研究人员会自行或与一小组研究助理一起尝试开展一个项目（例如肖文斯基最初所进行的分类）。如果这个方法在规模上不能满足要求，那么研究人员就会选择采用有许多参与者的人本计算。但在数据量达到一定规模后，单纯依靠人力就不足以解决问题了。在这种时候，研究人员就需要构建一个计算机辅助人本计算系统，利用人工分类数据训练一个基本能够处理无限数据的机器学习模型。

5.2.2 政治宣言的公众编码

通常由专家进行的政治宣言编码也可以通过人本计算来完成，后者还能实现更大的再现性和灵活性。

与星系动物园的发起人相似，社会研究人员在许多情况下也想对图像或文本进行编码、分类或标注。其中一个例子就是对政治宣言进

行编码。在竞选期间，政党会发表表明其政策立场和指导思想的宣言。例如，以下就是一则英国工党在2010年的宣言：

> 在我们的公共服务体系中工作的数百万名工作人员，他们践行着英国最崇高的价值观——为让人们能够充分利用他们的生命而贡献自己的力量并保护他们免受那些不应独自承受的风险的伤害。正如我们需要让政府在使市场公平运作方面扮演更加果敢的角色，我们也需要果敢地对政府进行改革。

对于政治科学家，尤其是那些研究竞选和政策辩论动态的政治科学家来说，这些宣言包含了非常有价值的数据。为了系统地从这些宣言中提取信息，研究人员创建了一个宣言项目，共搜集了50个国家的近1000个政党所发表的4000则宣言，然后组织政治科学家采用56类方案对每则宣言的每句话进行了系统的编码，最终建立了一个庞大的数据集，整合了这些宣言中所包含的信息。目前已有200多篇科学论文使用了该数据集。

肯尼思·贝努瓦（Kenneth Benoit）和同事决定把以前由专家进行的宣言编码转化成一个人本计算项目。于是，他们创建了一个更具再现性和灵活性的编码过程，其低成本和快速就更不在话下。

贝努瓦和同事采用分解－运用－整合这一策略，让来自微任务劳动力市场（机器人MTurk和众包公司CrowdFlower都是微任务劳动力市场的例子，更多有关该类市场的内容可参见第4章）的工人对英国从1987年到2010年间的6次大选所发表的18则宣言进行了编码。首先，研究人员将每则宣言分解为一个个句子。然后，工人运用编码方案对每句话进行编码。具体来说，他们被要求将每句话归类为经济政策（偏左或偏右）、社会政策（自由主义或保守主义）或两者都不是（图5.5）。每句话都大约有5个不同的人对其进行分类。最后，在考虑个人因素影响和句子难度影响的前提下，研究人员利用一个统计模型对

所有分类数据进行整合。最终贝努瓦和同事从大约1500名工人那里搜集了20万条分类数据。

图5.5 贝努瓦等人的编码方案。工人被要求将每句话归类为经济政策（偏左或偏右）、社会政策（自由主义或保守主义）或两者都不是。改编自Benoit et al.（2016），图1。

为了评估上述公众编码的质量，贝努瓦和同事还让大约10名专家，即政治科学领域的教授和研究生，用类似的步骤对相同的宣言进行了编码。尽管公众分类的个体一致性低于专家，但经过整合所得出的公众一致分类数据与专家一致分类数据吻合度非常高（图5.6）。与星系动物园一样，上述比较表明人本计算项目也能产生高质量的结果。

图5.6 在对英国政党发表的18则宣言进行编码时，公众编码评估结果与专家编码评估结果吻合度非常高。上述宣言是6次大选（1987年、1992年、1997年、2001年、2005年和2010年）期间三个政党（保守党、工党和自由民主党）所发表的。改编自Benoit et al.（2016），图3。

在此基础上，贝努瓦和同事利用他们的公众编码方法，开展了宣言项目专家无法完成的研究。例如，宣言项目的编码方案并没有涉及移民这一话题，因为在制订编码方案的20世纪80年代中期，移民并不是一个很突出的话题。但让宣言项目团队返回去重新编码他们的宣言以获取这一信息在组织实施上是不可行的。因此，有兴趣研究移民政治的研究人员似乎不太走运。但贝努瓦和同事利用他们的人本计算方法可以轻松且快速地进行这一编码。

为了研究移民政策，他们对英国2010年大选期间8个政党所发表的宣言进行了编码。每则宣言中的每个句子都需按照其是否与移民有关来编码，如果有关，还要判断是支持移民、中立，还是反对移民。项目启动后5个小时内，他们就搜集到了22000多条回复，总成本是360美元。而且，公众的评估结果与之前专家的评估结果吻合度非常高。两个月之后，他们又让公众对相同的宣言进行了一次编码，作为最后的

检测。然后，他们在几小时内便创建了一个新的与最初的公众编码数据集高度匹配的公众编码数据集。换句话说，人本计算使研究人员能够生成与专家评估一致的政治文本编码数据，而且该数据还具有再现性。此外，因为人本计算快速且便宜，所以他们可以很容易地根据移民政策这一特定的研究问题来调整他们的数据采集。

5.2.3 结论

人本计算能让你拥有1000名研究助理。

人本计算项目能够通过凝聚许多非专业人士的力量，解决那些计算机难以轻易解决的任务简单、数量庞大的问题。这类项目会采取分解–运用–整合这一策略将一个大问题分解成许多简单的、没有专业技能的人也能完成的微任务。计算机辅助人本计算系统还会利用机器学习放大人工成果的意义。

在社会研究中，当研究人员想要对图像、视频或文本进行分类、编码或标注时，最有可能用到人本计算。他们最终的目的通常并不是分类，而是在分类的基础上进行分析。例如，研究人员可以将对政治宣言公众编码数据的分析，作为对政治辩论动态这一更大课题的分析的一部分。效果最好的可能是不需要参与者接受专门的训练且参与者对任务的正确答案有着广泛共识的分类微任务。如果分类任务更具主观性，例如判断“这篇新闻报道有偏见吗”，那么了解参与者是谁以及他们的答案可能会有怎样的偏差将变得更加重要。最后，人本计算项目的输出质量取决于人工输入的质量：输入垃圾，则输出垃圾。

为了进一步增强你对人本计算的直觉，表5.1还列出了将人本计算用于社会研究的其他例子。该表格表明，与星系动物园不同的是，许多其他人本计算项目使用的是微任务劳动力市场（例如机器人MTurk平

台），即其依靠花钱雇人完成任务，而不是依靠志愿者来完成。当我提供有关如何创建自己的大规模协作项目的建议时，将再回到参与者激励这个话题。

表5.1 社会研究中的人本计算项目的例子

项目概述	数据类型	参与者	参考文献
政党宣言编码	文本	微任务劳动力市场	Benoit et al.（2016）
从有关美国200个城市的占领抗议的新闻文章中提取事件信息	文本	微任务劳动力市场	Adams（2016）
报纸文章分类	文本	微任务劳动力市场	Budak, Goel, and Rao（2016）
从第一次世界大战士兵的日记中提取事件信息	文本	志愿者	Grayson（2016）
发现地图中的变化	图像	微任务劳动力市场	Soeller et al.（2016）
检查算法编码	文本	微任务劳动力市场	Porter, Verdery, and Gaddis（2016）

最后，本小节的例子表明，人本计算让科学变得大众化了。肖文斯基和林托特开始创建星系动物园时都还只是研究生。在数字时代之前，一个对百万星系进行分类的项目应该需要花费大量时间和金钱，因此可能只有资金充足和有耐心的教授才能够开展。现在却不同了。人本计算项目通过凝聚许多非专业人士的力量，也能解决那些任务简单、数量庞大的问题。接下来我将向大家展示，大规模协作也适用于需要专业知识的问题，而这种专业知识有时甚至连研究人员自己也不具备。

5.3 公开征集

公开征集旨在为明确具体的目标征集新想法，它适用于“检验解决方案比想出解决方案更容易”的情形。

就上一小节所描述的人本计算问题来说，研究人员是知道如何解决这些问题的，只是没有足够的时间。也就是说，如果凯文·肖文斯基有无限的时间的话，他自己也能完成对100万个星系的分类。然而有时候，研究人员所面临的挑战不是来自规模方面，而是来自任务本身固有的难度。在过去，面临这类挑战的研究人员可能会向同事寻求帮助。而现在，这类问题还可以通过创建一个公开征集项目来解决。如果你曾想过“我不知道该如何解决这个问题，但我确信一定有人知道”，那么就可以通过公开征集解决这个问题。

在公开征集项目中，研究人员首先提出一个问题，向许多人征集解决方案，然后从解决方案中挑选最好的。把一个对自己来说很有挑战性的问题当作研究课题，然后借助公众来解决这个问题，这似乎有点奇怪，但我希望通过计算机科学、生物学和法律领域的三个例子，让大家相信这种方法很有效。这三个例子表明，创建一个成功的公开招募项目，关键在于仔细设计你的问题，尽管难以想出其解决方案，但你可以让该解决方案易于检验。然后在本小节的最后，我将更多地描述如何将这些想法应用于社会研究。

5.3.1 网飞奖

网飞奖通过公开征集来预测人们会喜欢哪部电影。

最著名的公开招募项目是网飞奖。网飞是一家在线电影租赁公司，它于2000年推出了电影匹配（Cinematch），一个向用户推荐电影的服务系统。例如，电影匹配可能注意到你喜欢《星球大战》和《帝国反击战》，然后据此向你推荐《绝地归来》。起初，电影匹配的表现差强人意。在过去的许多年里，它一直在不断提高预测用户喜好的能力。到2006年，电影匹配却停滞不前了。网飞的研究人员几乎尝试了所有他们能想到的东西，但同时他们猜想可能还有其他想法能帮助改进这个系统。于是，他们想到了一个在当时来说非常前卫的解决方案：公开征集。

对网飞奖最终的成功起到至关重要作用的是公开征集的设计，这个设计对公开征集如何才能被用于社会研究也有着重要的借鉴意义。网飞并没有像许多第一次接触公开征集这个概念的人设想的那样提出一个没有条理的征集想法，而是提出了一个明确且其解决方案易于检验的想法：要求人们利用1亿条电影评分数据来预测300万条评分留存数据（网飞没有公布的用户评分数据）。第一个开发出预测精度比电影匹配高10%的算法的人，将获得100万美元的奖金。而检验该算法的方法就是比较其预测评分与网飞的留存评分，这一明确且易行的检验方案意味着网飞奖的设计遵循了这样的理念：让检验解决方案比想出解决方案更容易。它将改进电影匹配的挑战变成了一个适于用公开征集来解决的问题。

2006年10月，网飞公开了一个数据集，其中包含了大约50万名用户的1亿条电影评分信息（我们将在第6章中讨论这一行为涉及的隐私问题）。这些数据可以被转化成一个巨大的矩阵，其中大约有50万名用户、2万部电影以及大约1亿条从1星到5星的电影评分信息（表5.2）。网飞的要求就是利用矩阵中的观测数据预测300万条留存评分。

表5.2 网飞奖数据简表

	电影 1	电影 2	电影 3	……	电影 20 000
用户 1	2 星	5 星		……	?
用户 2		2 星	?	……	3 星
用户 3		?	2 星	……	
……	……	……	……		……
用户 500 000	?		2 星	……	1 星

世界各地的研究人员和黑客都被这一挑战吸引了，到2008年，已有超过3万人参与其中（Thompson 2008）。在比赛过程中，网飞收到了来自5000多个团队的超过40000个解决方案提议（Netflix 2009）。显然，网飞无法阅读并理解所有这些方案。但整件事情进展很顺利，因为网飞很容易对解决方案进行验证。网飞只需让一台计算机按照预先指定的度量标准（他们当时采用的度量标准是均方误差的平方根）对预测评分和留存评分进行比较即可。正是这种快速评估解决方案的能力，使网飞能够评估每个团队的解决方案，而事实证明这一点很重要，因为好的创意确实来自一些令人惊讶的地方。事实上，获胜的解决方案来自一个由三位没有电影推荐系统构建经验的研究人员所组建的团队（Bell, Koren, and Volinsky 2010）。

网飞奖比较好的一点是所有方案都能得到公平的评估。也就是说，当人们上传预测评分数据时，无须上传学历、年龄、种族、性别、性取向或其他任何有关个人的信息。斯坦福大学一位著名教授的预测评分与一位青少年在其卧室中完成的预测评分所接受的评估是完全相同的。不幸的是，大多数社会研究却不是这样的。也就是说，对大多数社会研究来说，评估是非常耗时的，而且在一定程度上是具有主观性的。所以，大多数的研究想法从来都没有被认真评估过，而且在评估时，评估者也很难完全不考虑提出者身份这一因素。而公开征

集项目则有着公平易行的评估体系，所以它可以发现那些如果采用其他方法就会被忽略的想法。

例如，在比赛期间，有一个账号名为西蒙·芬克（Simon Funk）的人在他的博客上发布了一个基于奇异值分解的解决方案提议，这是一个线性代数的方法，其他参与者都未曾提到这一方法。芬克这篇博文既专业又很奇怪地不太正式。它描述的是一个好的解决方案，还是完全没用的东西？如果这不是一个公开征集项目的话，该解决方案可能永远也不会被认真评估。毕竟，西蒙·芬克并不是麻省理工学院的一位教授，他只是一名软件开发人员，当时正在新西兰背包旅行（Piatetsky 2007）。如果他当时通过邮件把这个想法发送给网飞的一位工程师，那么几乎可以肯定的是，这位工程师并不会认真评估这个想法。

幸运的是，因为网飞奖的评估标准很明确且评估易于实施，所以芬克的预测评分得到了评估，而且结果很快就出来了，他的方法显然非常有效：他的排名一下子飙升到了第4位。考虑到其他团队在这个问题上已经奋战了数月，这无疑是一个惊人的结果。最后，几乎所有认真对待这次比赛的竞争者都采用了他的部分方法（Bell, Koren, and Volinsky 2010）。

西蒙·芬克选择通过一篇博文来阐述他的方法，而不是试图避免让别人知道这个方法，这也表明网飞奖的许多参与者并不仅仅是因为百万美元的奖金才参与比赛的。更确切地说，许多参与者似乎也是为了享受这个问题所带来的智力挑战和由此而形成的团体（Thompson 2008），我觉得许多研究人员都能够理解这种感觉。

网飞奖是公开征集的一个经典示例。网飞提出了一个有着明确目标（预测电影评分）的问题，并向许多人征集该问题的解决方案。网飞之所以能够评估所有这些解决方案，是因为验证这些解决方案要比想出解决方案更加容易。最终，网飞经过验证挑选出了最好的解决方

案。接下来，我将向大家展示如何将同样的方法应用于生物学和法律领域，并且是在没有百万美元奖金的情况下。

5.3.2 蛋白质折叠游戏

蛋白质折叠游戏（Foldit）可以让非专业人士以一种有趣的方式参与进来。

网飞奖尽管很经典且明确易懂，但通过它并不能阐明公开征集项目的所有内容。例如，在网飞奖项目中，大多数认真对待比赛的参与者都受过多年统计学和机器学习方面的训练。但公开征集项目其实也可以让没有接受过正式训练的人参与，蛋白质折叠游戏就是这样。

蛋白质折叠是将氨基酸链折叠形成一定形状的过程。通过更好地理解这一过程，生物学家可以设计出具有特定形状的、能被用作药物的蛋白质。简单地说，蛋白质的折叠趋向于能量最低的构象，即蛋白质内部各种推力、拉力达到平衡状态时的一种结构（图5.7）。所以，如果一个研究人员想要预测蛋白质的折叠形状，其做法听起来其实很简单：只需尝试所有可能的构象并计算出每种构象的能量，然后预测蛋白质将折叠成能量最低的构象即可。不幸的是，尝试所有可能的构象从计算上来讲是不可能的，因为可能的构象有数十亿种。即便用现在最强大的计算机来做，在可预见的未来，这样的蛮力也不会起作用。因此，生物学家开发了许多巧妙的算法来有效地寻找最低能量的构象。但是，尽管在科学和计算方面付出了巨大的努力，这些算法还是远远不够完善。

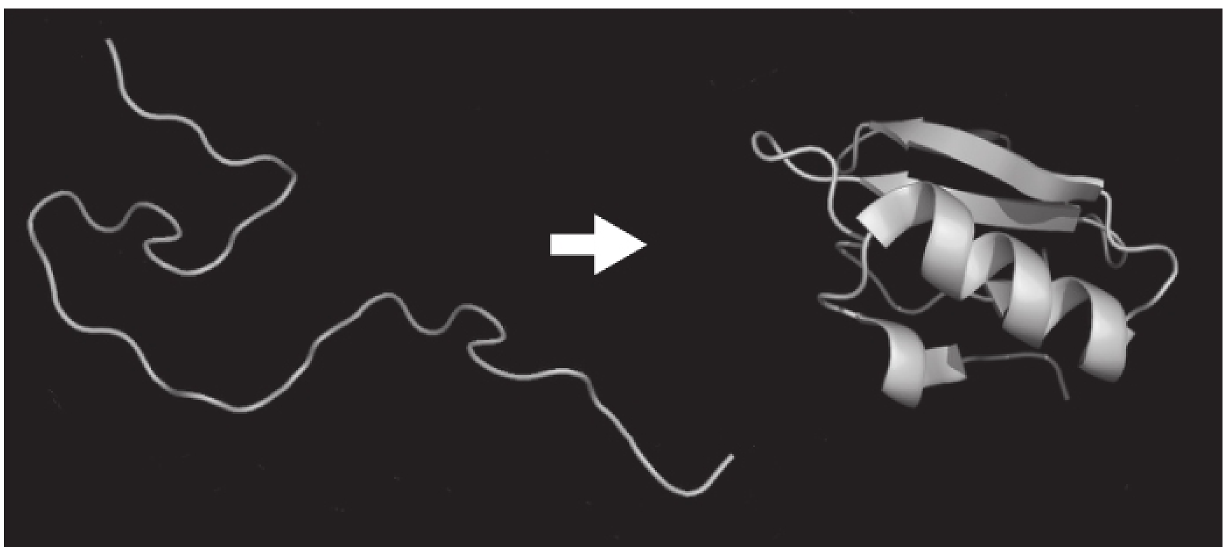


图5.7 蛋白质折叠。

华盛顿大学的戴维·贝克（David Baker）和他的研究小组同许多科学家一样致力于研究蛋白质折叠的计算方法。在一个项目中，贝克和同事开发了一个系统，志愿者可以利用空闲时间在他们的计算机上通过这一系统模拟蛋白质折叠。作为回报，他们所模拟的折叠方案可以成为他们计算机的屏幕保护图。然后，一些志愿者写信给贝克和他的同事说，他们认为如果自己能参与计算，就能提高计算机在预测蛋白质折叠方面的性能。于是蛋白质折叠游戏便诞生了（Hand 2010）。

蛋白质折叠游戏将蛋白质的折叠过程变成了一个任何人都可以玩的游戏。从玩家的角度来看，蛋白质折叠游戏就像是一个拼图游戏（图5.8）。玩家会看到一个杂乱无章的蛋白质三维结构图，然后通过“调整”、“扭转”和“重组”这些操作来改变它的形状，同时玩家的分数也会因此而增加或减少。重要的是，分数是根据当前构象的能量高低来计算的，能量越低，得分越高。换句话说，分数有助于指导玩家寻找低能量的构象。就像网飞奖的电影评分预测一样，蛋白质折叠游戏所呈现的也是一个“检验解决方案比想出解决方案更容易”的情形，也正因如此，这个游戏才得以诞生。

图5.8 蛋白质折叠游戏的游戏画面。

蛋白质折叠游戏的巧妙设计使那些对生物化学知之甚少的玩家也能与专家设计的最佳算法竞争。尽管大多数玩家都不是特别擅长，但也有少量个体玩家和小团队玩家例外。事实上，在玩家与最先进的算法的正面交锋中，玩家的折叠方案更胜一筹的概率是50%（Cooper et al. 2010）。

蛋白质折叠游戏和网飞奖在许多方面都不一样，但它们都公开征集解决方案，而且都是“检验解决方案比想出解决方案更容易”。现在，我们将在另一个非常不同的领域看到相同的结构：专利法。最后这个公开征集的例子表明，这一方法也适用于看起来不是很易于量化的情形。

5.3.3 公众专利评审

公众专利评审是一项帮助专利审查员发现现有技术的公开征集项目。它表明公开征集也可以被用于不易量化的问题。

专利审查员的工作是很辛苦的。他们会收到关于新发明的、简洁的、如律师般措辞严谨的描述文件，然后必须决定该发明是否“新颖”。也就是说，审查员必须审查是否存在会使所陈述的发明无法获得专利的“现有技术”（之前已有对所陈述发明的描述文件）。我们可以通过一个名叫阿尔伯特（为了纪念阿尔伯特·爱因斯坦）的专利审查员来理解这个过程是如何运作的。阿尔伯特的职业生涯开始于瑞士专利局，他可能会收到一份类似于美国专利20070118658的申请。美国专利20070118658是惠普为其“用户可选管理警报格式”所申请的，贝丝·诺维克（Beth Noveck）的《维基政府》（*Wiki Government*）中有关于这项专利的大量描述。以下是这个专利申请的第一条权利要求。

一个计算机系统包括：一个处理器、一个包括逻辑指令的基本输入/输出系统（BIOS）。当处理器执行该逻辑指令时，它便会对处理器进行如下配置：在一个计算设备的基本输入/输出系统中启动开机自检（POST），在用户界面中呈现一个或多个管理警报格式，从用户界面接收选择信号，识别用户界面中所显示的其中一个管理警报格式，以及用已识别的管理警报格式配置一个与计算系统相连的设备。

阿尔伯特是否应该授予这项申请20年的垄断权呢？是否存在会使它无法获得专利的现有技术？许多专利的利害关系很大，但不幸的是，阿尔伯特不得不在没有太多他可能需要的信息的情况下做出这个决定。因为有大量的专利申请积压，所以阿尔伯特一直在巨大的时间压力下工作，审查时间只有20个小时，然后他必须做出决定。此外，由于正在审查的发明需要保密，所以这项工作也不允许阿尔伯特咨询外部专家（Noveck 2006）。

这一状况让法学教授贝丝·诺维克感到很困惑。2005年7月，在一定程度上受维基百科的启发，她写了一篇标题为《公众专利评审：一个小的提议》的博文，呼吁建立一个公众专利评审制度。通过与美国专利商标局以及IBM（国际商业机器公司）等领先科技公司合作，公众专利评审于2007年6月正式启动了。一个是有将近200年历史的政府官僚机构，一个是律师群体，它们似乎都不太可能寻求创新，但公众专利评审巧妙地为每个人找到了一个利益平衡点。

以下是公众专利评审的工作流程（图5.9）。在发明者同意对他的申请进行公开审查后（稍后我会分析这个人这么做的可能原因），其申请就会被上传到一个网站。然后，由公众审查人员（稍后我会分析他们这么做的可能原因）组成的审查小组对这一申请进行讨论，并查找、标注和上传与之相似的现有技术。这一过程会一直进行，直到审查小组最终投票选出最有可能与之相似的10个现有技术为止，然后他

们会将这10个现有技术发送给专利审查员进行审查。专利审查员再自行进行审查，然后结合来自公众专利评审的意见做出最终判断。

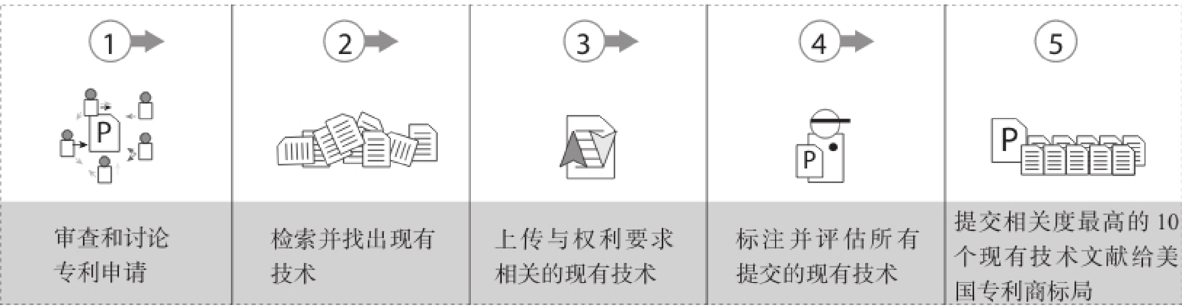


图5.9 公众专利评审的工作流程。复制自Bestor and Hamp（2010）。

让我们再次回到有关“用户可选管理警报格式”的美国专利20070118658这个话题。这个专利是在2007年6月被上传至公众专利评审的，然后IBM的高级软件工程师史蒂夫·皮尔逊（Steve Pearson）看到了这个专利的内容。皮尔逊对这一研究领域很熟悉，并找到了一个现有技术文献：一本标题为《主动管理技术：快速参考指南》（*Active Management Technology: Quick Reference Guide*）的英特尔的指南手册，这本手册在两年前就出版了。在这份文件以及其他现有技术文献和公众专利评审中审查小组讨论的基础上，一名专利审查员开始对该专利进行彻底的审查，并最终撤销了该项技术的专利权，部分原因就是皮尔逊发现的英特尔的指南手册（Noveck 2009）。在通过公众专利评审所完成的66个专利申请中，有近30%主要是因为评审小组发现的现有技术而被拒绝授予专利权的（Bestor and Hamp 2010）。

公众专利评审设计的巧妙之处，在于它为有许多利益冲突的相关方提供了一个和谐协作的平台。发明者之所以参与是因为通过公众专利评审所提交的申请，要比走传统的秘密审查程序的申请更快获得专利局的审查。公众审查人员之所以参与是为了防止低质量专利产生，而且许多人似乎觉得这个过程很有趣。最后，专利局和专利审查员之所以参与是因为这个方法只会改善他们的审查结果。也就是说，如果

评审小组发现的是10个无用的现有技术文献，那么专利审查员将它们忽略掉即可。换句话说，有公众审查人员与专利审查员一起合作，应该比专利审查员独自奋战要好，这至少也应该能达到与专利审查员独自奋战相同的效果。因此，公开征集并不总是代替专家解决问题，有时是帮助专家把他们的工作做得更好。

尽管公众专利评审与网飞奖和蛋白质折叠游戏不太一样，但它们都有一个相似的结构，即“检验解决方案比想出解决方案更容易”。一旦有人找到了《主动管理技术：快速参考指南》这本手册，那么核实这个文件是否是现有技术就容易了，至少对专利审查员来说是这样的。然而发现这本手册是相当困难的。公众专利评审还表明，公开征集有时也适用于不是很易于量化的问题。

5.3.4 结论

公开征集可以让你为那些你能清楚地描述但无法自己解决的问题找到解决方案。

在上述三个公开征集的项目，即网飞奖、蛋白质折叠游戏和公众专利评审中，研究人员都是先提出一个特定形式的问题，然后公开征集解决方案，最后挑选最好的解决方案。研究人员甚至都不需要知道可以请教的最好的专家是谁，其实好的想法有时会来自意想不到的地方。

鉴于我已经对人本计算项目和公开征集项目进行了介绍，所以现在我可以强调两者之间的两个重要区别了。首先，在公开征集项目中，研究人员指定的是一个目标（例如预测电影评分），而在人本计算项目中，研究人员指定的是一个微任务（例如对一个星系进行分类）。其次，在公开征集项目中，研究人员想要的是最好的贡献，例

如预测电影评分的最好算法、蛋白质的最低能量构象或者最相关的现有技术文献，而不是所有贡献的简单集合。

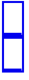
我已经描述了公开征集的通用模板以及三个示例，那么社会研究中的哪些问题适合用该方法来解决呢？在这一点上，我得承认成功的例子还不是很多（我稍后会解释原因）。就直接模拟而言，我们可以设想一下，一位历史研究人员通过公众专利评审式的公开征集查找最早提及某个特定的人或想法的文件，尤其是当潜在的相关文件没有被归档在一起，而是广泛分散在各处时，公开征集对这类问题的价值就更大了。

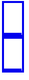
更普遍地讲，许多政府和公司都有需要公开征集解决的问题，因为公开征集能够产生预测算法，而这些预测可以成为行动的重要指南（Provost and Fawcett 2013; Kleinberg et al. 2015）。例如，就像网飞想要预测电影评分一样，政府可能想要预测哪家餐馆最有可能违犯卫生法规等，以便更有效地分配检查资源。鉴于此，爱德华·格莱泽（Edward Glaeser）和同事便利用公开征集帮助波士顿市根据点评网站Yelp上的点评数据和历史检查数据预测餐馆的卫生违规情况。他们估计，通过公开征集所选出的最好的预测模型能够将餐馆检查员的工作效率提高约50%。

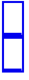
公开征集还可能被用于比较和测试理论。例如，脆弱家庭和儿童福利研究（Fragile Families and Child Wellbeing Study）对在美国20个不同城市出生的约5000个孩子进行了追踪（Reichman et al. 2001）。研究人员分别在孩子出生时以及1岁、3岁、5岁、9岁和15岁时搜集了有关这些孩子、他们的家庭以及他们所处的更广泛的环境方面的数据。那么研究人员利用所有这些数据预测谁将能够大学毕业的准确度将如何呢？或者用有些研究人员更加感兴趣的方式来表达就是，哪些数据和理论在预测这些结果方面最有效？因为这些孩子目前都还没到上大学的年龄，所以这将是一个真正前瞻性的预测，而且研

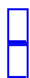
究人员可能采取的策略也有许多。认为社区对大学毕业等结果至关重要的研究人员与认为家庭至关重要的研究人员所采用的方法可能截然不同。那么哪种方法会更有效呢？我们并不知道，在寻找答案的过程中，我们可能会学到一些有关家庭、社区、教育和社会不平等的重要知识。此外，这些预测也许能被用来指导以后的数据采集。假如有一小部分大学毕业生在之前各种方法的预测中都是不可能大学毕业的，那么这些人将是后续定性采访和人种志观察的理想人选。因此，在这一类型的公开征集中，预测并不是目的，相反，它们为比较、拓展和结合不同的理论传统提供了一种新方法。这种公开征集不只适用于利用脆弱家庭和儿童福利研究的数据来预测谁将上大学，它还可以预测最终会被搜集到纵向社会数据集中去的所有结果。

正如我在本小节前面部分所写的，目前还没有很多社会研究人员采用公开征集方法的例子。我认为这是因为公开征集的提问方式与社会科学家通常的提问方式不太一样。社会科学家通常不会问有关预测品位的问题，相反，他们会问不同社会阶层的人的文化品位为什么会不同以及是如何不同的（参阅例如Bourdieu 1987）。像这种“如何”以及“为什么”的问题，其解决方案通常都不易于检验，因此似乎不适用于公开征集。所以，似乎公开征集更适用于预测性问题而不是解释性问题。但最近理论家已经开始呼吁社会科学家重新考虑解释和预测之间的分界线（Watts 2014）。随着解释和预测之间的界线越来越模糊，我希望公开征集在社会研究中能变得越来越普遍。









第6章　道德伦理

6.1 简介

在之前的章节中，我已经展示了数字时代在搜集和分析社会资源方面所创造的新机遇。与此同时，数字时代也带来了新的道德伦理挑战。本章的目的在于为你们提供相应的工具，用以负责地处理这些道德伦理的挑战。

当下，一些数字时代社会研究的道德伦理问题还存在着不确定性。这种不确定性导致两种相关问题产生，其中一个比另一个受到了更多的关注。一方面，一些研究人员被指控侵犯了人们的隐私权或者参与了不道德的实验。我即将在本章描述的事例，已经引起广泛争议并成为讨论的事实主体。另一方面，道德伦理上的不确定性也产生了令人不寒而栗的结果，这些结果阻碍了道德伦理以及一些重要方面的研究，这一事实我认为还不太受人重视。譬如，在2014年埃博拉疫情暴发期间，公共卫生当局希望获取在疫情最严重的相应国家的人群迁移信息，从而帮助控制疾病的传播。移动通信公司拥有详细的通信记录，并可以从中获取相关的信息。然而，道德伦理以及法律方面的担忧使研究人员分析数据的尝试陷入困境（Wesolowski et al. 2014; McDonald 2016）。如果我们作为社群的一员，能够制定研究人员和公众共享的道德伦理规范和标准（这一点我认为可以做到），那么我们就可以用一种对社会负责和有益的方式来利用数字时代赋予我们的能力。

一个阻碍制定该共享标准的因素在于，社会科学家与数据科学家倾向于采用不同的方法研究道德伦理。对社会科学家而言，道德伦理的思考由机构审查委员会主导，其任务在于执行一系列的法规。毕竟，对大多数实证社会科学家来说，经历道德伦理辩论的唯一途径是

通过官僚机构审查委员会的审查过程。数据科学家对研究道德伦理问题仅有一些少量的系统性经验，因为对计算机科学与工程学来说，这些问题并不是受到普遍讨论的。无论是社会科学家在乎的以规则为基准的方法，还是数据科学家在乎的以特例假设为基准的方法，两者在数字时代对于社会研究均不适用。取而代之，我相信我们作为社群的一员，如果能够采取一种基于原则的方法，就能取得进步。也就是说，研究人员应该用现有的规则来评估他们的研究（如果有的话），以及用更为普遍的道德伦理原则进行评估。这种基于原则的方法能帮助研究人员在规则不适用的情况下做出理性的决定，并且帮助研究人员将他们的推断传达给他人和公众。

我所提倡的这种基于原则的方法并非初创。它借鉴了数十年前的一些想法，其中大部分内容都体现在两个具有里程碑意义的报告中：《贝尔蒙报告》（*Belmont Report*）与《门罗报告》（*Menlo Report*）。你将看到，在一些事例中，这种基于原则的方法能引领我们找到清晰有效的解决方法。如果不能解决这些问题，它将阐明所涉及的需要权衡的问题，这对于实现适当的平衡至关重要。此外，基于原则的方法具有足够的共通性，无论你们在哪里工作（比如大学、政府机关、非政府组织或者公司），它都会有所帮助。

本章旨在帮助善意的个体研究人员。对于工作中所涉及的道德伦理问题，你应该怎样思考？你应该如何让你的工作更加符合道德伦理？在6.2节，我将介绍三个在数字时代引起道德伦理争议的研究项目。接着在6.3节中，我会将这些具体的事例抽象化，以此描述产生道德伦理不确定性的一些基本原因：迅速提高的研究人员相关能力，使得他们在未经参与者同意甚至在其毫不知情的情况下对人类进行观察与实验。这种能力的改变远超我们的规范、准则与法律的修订速度。在6.4节中，我将描述四项既有的原则，以便于指导你的思考方向：对他人的尊重原则、有利化原则、公正原则以及对法律和公共利益的尊重原则。接下来，在6.5节中，我会总结两种广泛的道德框架，即结果

主义与义务论，它们可以帮助你解决可能面临的更深层次的挑战：在什么样的时机适合使用道德上存在问题的手段来达到符合道德标准的目的？这些原则和道德框架在图6.1中得以归纳，它们将使你超越对现有法规所允许的范围的关注，并提高你与其他研究人员和公众沟通想法的能力。

图6.1 支配研究的规则来源于原则，而原则相应地来源于道德框架。本章的一个主要论点在于研究人员应该通过现有的规则（如果有的话）以及更为一般的道德原则评估自己的研究。《通则》（即《美国联邦受试者保护通则》）是目前支配美国大部分联邦资助研究的一系列规定（更多信息参见本章的历史附录）。这四项原则来自《贝尔蒙报告》和《门罗报告》（更多信息参见本章的历史附录）。最后，结果主义与义务论作为两种道德框架已经在过去的数百年中经过哲学家的努力得以发展。有一个简单而粗略的方法，可以用来区分这两种框架：义务论专注于过程，结果主义专注于结果。

有了这样的背景，在6.6节中，我将讨论对数字时代的社会研究人员来说，特别具有挑战性的四个领域：知情同意（6.6.1小节）、理解与管理信息化风险（6.6.2小节）、隐私权（6.6.3小节）以及面对不确定性做出符合伦理规范的决策（6.6.4小节）。在6.7节，我将提供在不稳定的伦理领域工作环境下的三个实用技巧。本章最后是一个历史附录，我简要总结了美国道德伦理监督研究的进展，包括对塔斯基吉梅毒实验（*Tuskegee Syphilis Study*）、《贝尔蒙报告》、《通则》和《门罗报告》的讨论。

6.2 三个事例

数字时代的社会研究将涉及一些理性而善意的人不会同意的道德伦理情况。

为了让事情具体化，我将从三个引起道德伦理争议的数字时代研究项目开始讲起。我选择这些特别的研究项目主要基于两个理由。一是关于这些研究均没有简单的答案。也就是说，理性而善意的人们在这些研究是否应该发生以及哪些改变可能改善这些研究项目方面意见不一致。二是这些研究体现了本章后半部分将要讨论的许多原则、框架及其紧张关系。

6.2.1 情绪感染项目

70万名脸谱网用户被放入一项实验中，这可能改变他们的情绪。这些参与者并没有同意参与，该研究也没有受到有意义的第三方伦理监管。

2012年1月的一周里，大约70万名脸谱网用户被安置进一个名为情绪感染的实验中，即研究一个人的情绪受到与其互动的他人情绪影响程度的实验。我已经在第4章中讨论过这个实验，但是我现在还会再次回顾它。情绪感染这项实验中的参与者被放入4个组：“消极减少”组，即在新消息展示中对那些带有消极词汇（比如“悲伤”）的帖子进行随机屏蔽；“积极减少”组，即对那些带有积极词汇（比如“幸福”）的帖子进行随机屏蔽；以及两个对照组，一个对应“积极减少”组，另一个对应“消极减少”组。研究人员发现，与对照组相

比，“积极减少”组的人使用的积极词汇略少，而消极词汇略多。相似地，他们也发现身处“消极减少”组的人使用积极词汇稍多，消极词汇较少。因此，研究人员得出了情绪感染的相应证据（Kramer, Guillory, and Hancock 2014）。更完整的实验设计和结果的讨论，请参见第4章。

在该论文通过《美国国家科学院院刊》发表后，研究人员和新闻媒体对此表达了强烈的抗议。围绕该论文的争议主要集中在以下两点：（1）对于超出脸谱网标准服务条款规则的部分，参与者并没有提供任何的同意许可；（2）该研究没有经过有效的第三方伦理审查（Grimmelmann 2015）。日渐激化的争议中所反映的道德伦理问题，使该期刊迅速针对此次研究发表了一篇罕见的关于道德伦理和伦理审查过程的社论（Verma 2014）。在随后的几年中，这项实验一直广受争议并引起了激烈的讨论，而对这项实验的批评可能会将这类研究引向地下实验，造成意想不到的后果（Meyer 2014）。也就是说，有些人认为某些公司并没有停止这类实验，只是停止了在公众面前提及它们。无论如何，这场争议可能有助于促进脸谱网伦理研究审查流程的创建（Hernandez and Seetharaman 2016; Jackman and Kanerva 2016）。

6.2.2 “3T”项目

研究人员从脸谱网上搜刮学生的数据，将其与大学记录合并，将这些合并后的数据用于研究，然后与其他研究人员共享。

从2006年开始的每一年，一个由教授及其研究助理组成的小组都在“美国东北部的一所多元化私立大学”中搜刮学生的脸谱网资料。然后研究人员将这些包含了好友信息与文化品位的脸谱网数据与包含了学术主攻方向和在校园内居住信息的校方数据进行了合并。这些合

并后的数据是非常有用的资源，通过这些数据，我们能够对社交网络是如何形成的（Wimmer and Lewis 2010）以及社交网络与行为是如何共同演变的（Lewis，Gonzalez and Kaufman 2012）等主题产生新的认识。除了将这些数据用于他们自己的工作之外，“3T”项目［Tastes（文化品位）、Ties（关系）、Time（时间）］的研究人员在采取一些保护学生隐私的措施后，还将这些数据提供给了其他研究人员（Lewis et al. 2008）。

不幸的是，在提供这些数据的数日后，其他研究人员就推断出这所学校是哈佛大学（Zimmer 2010）。这个项目的研究人员被指责为“不遵守伦理研究标准”，部分原因在于学生并未提供知情同意（所有程序均经过哈佛机构审查委员会和脸谱网审核并批准）。此外还出现了来自学术机构、媒体的批评声，比如标题为《哈佛研究人员被控侵犯学生隐私权》的纸媒报道（Parry 2011）。最后，这些数据库被从网上移除，并且不再能提供给其他研究人员使用。

6.2.3 “Encore”项目

研究人员让人们的计算机秘密地去访问可能被专制政府屏蔽的网站。

2014年3月，山姆·伯内特（Sam Burnett）与尼克·菲姆斯特（Nick Feamster）启动了“Encore”（意为“再次”）项目，这是一个为互联网审查提供实时和全球测量的系统。要做到这一点，位于乔治亚理工学院的研究人员鼓励网站所有者将这个小代码片段放到他们网页的源文件中：

```
<iframe src="//encore.noise.gatech.edu/task.html"
        width="0" height="0"
        style="display: none"></iframe>
```

如果你访问一个植入此片段代码的网站主页，你的网页浏览器就会尝试与网站进行沟通，以使研究人员发现可能的互联网审查（比如观察你是否访问了一个被禁止的政治党派的网站）。接着，你的网页浏览器将会反馈给研究人员，告诉他们它是否能够与可能受到屏蔽的网站进行沟通（如图6.2）。更进一步来说，所有的步骤都不可见，除非你调用网页的源文件进行检查。这种隐形的第三方页面请求在网络上非常普遍（Narayanan and Zevenbergen 2015），但它们很少涉及明确的对互联网审查的测量。

这种测量互联网审查制度的方法有一些非常具有吸引力的技术特性。如果足够多的网站都植入了这样简单的代码片段，那么“Encore”项目就能够提供对被审查网站的实时化、全球化测量结果。在启动这个项目之前，研究人员与所处机构的机构审查委员会进行了交流，而该机构审查委员会拒绝审查该项目，因为它不满足《通则》（更多信息请参阅本章历史附录）规定下的“人体研究”条款。

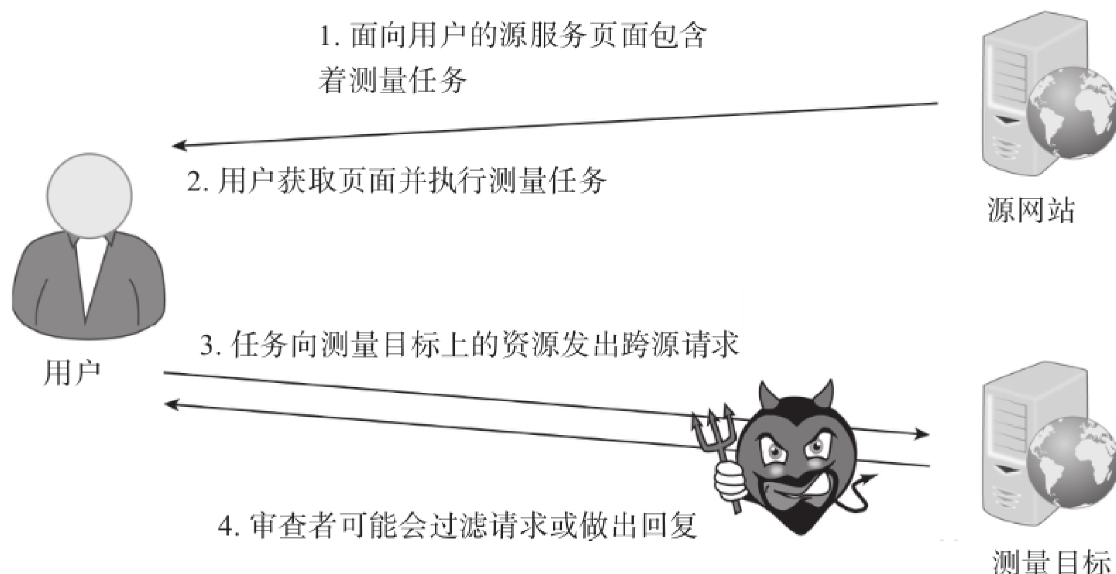


图6.2 “Encore”项目的研究设计示意图。在源网站中植入一小段代码片段（步骤1）。你的计算机加载该网页时，随之启动测量任务（步骤2）。你的计算机试图访问该测量目标，也就是访问受到屏蔽的政党网站（步骤3）。政府审查互联网的审查者在接下来可能会对你向测量目标的访问进行屏蔽（步骤4）。最终你的计算机会反馈给研究人员相应的请求结果（在图中未显示）。该图片获得计算机协会Burnett and Feamster（2015）文中图1的转载许可。

然而，在“Encore”项目启动后不久，研究生本·泽文贝根（Ben Zevenbergen）联系了该项目的研究人员，提出了有关“Encore”项目涉及道德伦理的问题。特别是，泽文贝根关注那些在特定国家里利用他们的计算机尝试访问某些敏感网站的人们，他们可能会被暴露在危险情况之下，并且这些人在参与这个研究的过程中并不知情。基于这些对话，“Encore”项目团队修改了该项目的运营条件，试图仅对脸谱网、推特以及优兔进行测试，因为在正常的网页浏览器中，第三方试图访问这些网站是很常见的（Narayanan and Zevenbergen 2015）。

在使用这种修改后的设计搜集数据后，一份描述其算法和一些结果的论文被提交给了一个著名的计算机科学会议，即美国计算机学会数据通信专业组（SIGCOMM）。该项目委员会对该文献的技术贡献表示赞赏，但对其缺乏参与者知情同意表示担忧。随后，该项目委员会决

定发表该论文，但也随之附上一份对道德伦理表示关注的声明（Burnett and Feamster 2015）。这种类似的附属声明从未在数据通信专业组会议上被使用过，这个事例引发了计算机科学家对他们的研究中有关伦理性质问题的讨论（Narayanan and Zevenbergen 2015; Jones and Feamster 2015）。

6.3 数字时代的不同

在数字时代进行的社会研究有许多不同的特性，这些特性带来了不同的道德伦理问题。

在模拟时代，绝大多数社会研究的规模相对有限，它们在一系列合理而明确的规则下运作。但在数字时代，社会研究的情况截然不同。研究人员通常与公司或政府部门进行紧密的合作，相比于过去，他们对测试参与者拥有了更多的掌控力，而对这些力量的使用并没有一个明确的规则。对于这种能力，我将其简单地视作在未经人们知情同意甚至在其没有意识的情况下进行研究活动的能力。这一系列的活动包括研究人员能够在实验中观察人们并控制他们的行为。随着研究人员观察和干预能力的增强，对如何使用这种能力的清晰定义却并没有被相应界定。事实上，研究人员必须基于一些前后矛盾并相互重叠的规则、法律和规范来决定如何行使这些能力。强大的能力与模糊的指导方针的结合造成了当下困难的情况。

这样一系列能力中的一项，包括研究人员现在可以未经参与者知情同意或者在他们没有察觉的情况下观察参与者的行为。研究人员在过去当然也可以这样做，但是在数字时代，这种规模是完全不同的，这一事实已经被许多大数据热衷者反复宣传过了。特别是，如果我们从个别学生或教授的研究规模转移到公司或者政府部门层面上，比如对与研究人员合作日益紧密的机构来说，潜在的伦理问题就变得复杂起来。我认为有一个比喻可以帮助人们全面了解这种大规模监视的想法，那就是全景监狱（Panopticon）。全景监狱是最初由杰里米·边沁（Jeremy Bentham）针对监狱而提出的一种建筑设计，它是一种圆形建筑，监舍是围绕着中央瞭望塔而修建的（图6.3）。监狱管理者能

够在瞭望塔中观察所有在监囚犯的行为，而监舍中的犯人无法看到监视人员。身处瞭望塔中的人们因此被称为看不见的观察者（Foucault 1995）。对于一些隐私倡导者来说，数字时代将我们带进了全景监狱，因为科技公司与政府部门不断监视并记录我们的行为。

图6.3 全景监狱由杰里米·边沁首次提出。居中处有一名看不见的观察者能够监视所有人的行为，却不会被他人察觉。上图由威利·雷瓦尔利（Willey Reveley）于1791年绘制。来源：《杰里米·边沁的工作》（*The Works of Jeremy Bentham*）一书。

由此比喻，我们发散开来，许多社会研究人员认为，在数字时代，他们可以想象自己是身处中心瞭望塔中的一员，观察人们的行为并创建一个主数据库，进行各种各样让人激动的重要研究。但是现在，不妨设想你不再是身处中心瞭望塔中，而是身处其中一个监舍之中，那么这个主数据库就开始被视作堕落数据库［由保罗·欧姆（Paul Ohm）在2010年提出］，可以用于不道德的途径。

本书的一些读者足够幸运，能生活在一个他们信任其看不见的观察者能够负责任地使用这些数据，并且面对敌人能够保护数据的国家里。另外一些读者却不够幸运，并且我确定大规模监控所带来的问题对他们来说是非常明确的。但是我相信对那些幸运的读者来说，大规模监控仍旧会带来一个非常值得关注的问题：意料之外的二次使用。也就是说，一个数据库是出于某一种目的而建造的，比如发送定向广告，但有一天它也可能被用于一种截然不同的目的。一个让人毛骨悚然的事例，也是意料之外的二次使用，发生在第二次世界大战期间，当时政府的人口普查数据被用来促进对犹太人、罗姆人和其他一些人种的种族灭绝（Seltzer and Anderson 2008）。这些在和平时期搜集数据的统计学家几乎都明确地拥有良好的动机，并且大多数人都相信他们能够负责任地使用这些数据。但是，当世道改变的时候，纳粹党获得权力，这些数据都出乎意料地被二次使用。原因很简单，一旦存在主数据库，就很难意料到谁可以访问它，以及如何使用它。事实

上，威廉·塞尔策（William Seltzer）与马戈·安德森（Margo Anderson）2008年就记录了18项人口统计数据涉及或可能涉及侵犯人权的案件（表6.1）。更进一步来说，正如塞尔策与安德森提出的，这份列表几乎可以肯定只是保守估计，因为大多数数据滥用都是秘密发生的。

普通的社会研究与通过二次使用侵犯人权的行为相去甚远。尽管如此，我选择讨论这项议题，是因为我认为它有助于你们理解一些人可能对你的工作作何反应。让我们回到“3T”项目作为事例。通过将来自脸谱网的完整精细的数据与哈佛大学的完整精细的数据合并在一起，研究人员对学生的社会和文化生活形成了惊人的丰富观点（Lewis et al. 2008）。对许多社会研究人员来说，这就像是主数据库，可以为良好的目的而服务。但是对其他一些人来说，这也可以让它成为一个堕落数据库，为不道德的目的而服务。事实上，可能是两者并存。

除了大规模监控，研究人员通过再次与公司和政府部门合作，可以越来越多地干预人们的生活，创建随机对照实验。比如，在情绪感染项目中，研究人员在未取得知情同意且参与者毫不知情的情况下，在实验中对70万人进行了控制。正如我在第4章所描述的那样，在这种实验中，秘密被征用的参与者并不少见，并且不需要大公司的合作。事实上，在第4章中，我已经指出了如何应对这样的情况。

表6.1 人口数据系统涉及或可能涉及侵犯人权的案例

地区	时间	目标个体或群体	数据系统	侵犯人权或推测为国家意图
澳大利亚	19世纪到20世纪初	原住民	人口登记	强制迁徙、种族灭绝
法国	1940—1944年	犹太人	人口登记、特殊人口普查	强制迁徙、种族灭绝
德国	1933—1945年	犹太人、罗姆人以及其他人种	众多措施	强制迁徙、种族灭绝
匈牙利	1945—1946年	德国国籍以及以德语为母语的人群	1941年人口普查	强制迁徙
荷兰	1940—1944年	犹太人和罗姆人	人口登记系统	强制迁徙、种族灭绝
挪威	1845—1930年	萨米人和克文人	人口普查	种族清洗
挪威	1939—1944年	犹太人	特殊人口普查和目的性人口登记	种族灭绝
波兰	1939—1943年	犹太人	特殊人口普查	种族灭绝
罗马尼亚	1941—1943年	犹太人和罗姆人	1941年人口普查	强制迁徙、种族灭绝
卢旺达	1994年	图西族	人口登记	种族灭绝
南非	1950—1993年	非洲人以及有色人种	1951年人口普查及人口登记	种族隔离、剥夺公民选举权
美国	19世纪	美洲原住民	特殊人口普查、人口登记	强制迁徙
美国	1917年	涉嫌违反法律草案者	1910年人口普查	对拒绝登记的人进行调查和起诉
美国	1941—1945年	日裔美国人	1940年人口普查	强制迁徙及拘留
美国	2001—2008年	疑似恐怖分子	国家教育统计中心调查及行政数据	国内和国际恐怖分子的调查与起诉
美国	2003年	阿拉伯裔美国人	2000年人口普查	未知
俄国及苏联	1919—1939年	少数民族	多种人口普查	强制迁徙、其他严重的惩罚

注：此表基于塞尔策与安德森2008年的报告所做，其中我列了一列子项目。有关每个事例和纳入标准的更多信息，请参见Seltzer and Anderson（2008）。其中一些事例涉及数据意料之外的二次使用，但并非全部如此。

面对这种日益增强的能力，研究人员受制于不统一和相互重叠的规则、法律和规范。这种不统一性的一个来源就是数字时代的能力发展远超过规则、法律和规范的修订速度。譬如，《通则》从1981年以来就没有大幅修订过，并且最近的一项试图使其更符合现代要求的提议需要近5年半的时间才能完成（Jaschik 2017）。另一个不统一性的来源就是，围绕着隐私权等抽象概念的规范仍在被研究人员、政策制定者和其他积极分子激烈讨论。如果在这一领域的专家都不能够达成一种统一意见，那么我们就不应该期望实证研究人员或参与者去做这些。第三个，也是最后一个不统一性的来源在于，数字时代的研究与其他领域或环境的联系更为紧密，这导致一些规范和规则存在潜在的重叠。譬如，在情绪感染项目中，来自脸谱网的数据科学家与来自康

奈尔大学的教授和研究生进行紧密合作。在那时，脸谱网进行了大量实验却没有第三方监管的事实是普遍的，只要这些实验符合脸谱网的服务条款即可。对康奈尔大学来说，规范与规则就截然不同：几乎所有的实验都必须在康奈尔大学机构审查委员会的监管下开展。那么，究竟应该采取什么样的规则来指导情绪感染项目，是听从脸谱网的还是康奈尔大学的？当这些都不统一并且与规则、法律和规范相互重叠时，充满善意的研究人员也可能遇到麻烦。事实上，正因为这样的不统一性，可能没有一件事是完全正确的。

总的来说，这两项特征，即能力的增强和应该如何使用能力的共识的缺乏，都意味着数字时代的研究人员将在可预见的未来面临道德伦理的挑战。幸运的是，在处理这些挑战时，并不需要从头开始。相反，研究人员可以从以前制定的道德原则和框架中吸取智慧。我将在下面两节中讨论这些主题。

6.4 四项原则

研究人员面对道德伦理不确定性时，可采取的四项原则包括：对人的尊重原则、有利化原则、公正原则、对法律和公共利益的尊重原则。

数字时代研究人员所面对的道德伦理挑战与以往截然不同。尽管如此，研究人员还是可以通过早期构建的伦理思想来应对这些挑战。特别是我相信《贝尔蒙报告》与《门罗报告》这两份报告反映出的原则可以帮助研究人员推断出他们面临的道德挑战。正如我在本章的历史附录中详细描述的那样，这两份报告都是多名专家组成的小组经过多年审议得出的结果，各种利益相关者提供了许多参考要素。

首先，在1974年，为回应研究人员的伦理过失（比如臭名昭著的塔斯基吉梅毒实验，在这个实验里，近4万名非洲裔美国男性被研究人员哄骗，并且在近40年的时间里无法获得安全而有效的治疗），美国国会设立了一个全国委员会，制定涉及人体研究的伦理准则。在贝尔蒙会议中心召开会议的4年后，该委员会制定了《贝尔蒙报告》，这是一份页数不多却分量十足的文件。《贝尔蒙报告》是《通则》的知识基础，而《通则》是由机构审查委员会强制执行的一系列用以指导以人类为对象的科学研究的规章制度（Porter and Koski 2008）。

接下来，在2010年，为了回应计算机安全领域研究人员的伦理过失，以及弥补在数字时代《贝尔蒙报告》观点应用的局限性，美国政府，特别是美国国土安全部，设立了一个蓝带委员会，为涉及信息通信技术的研究制定了一个指导性的道德框架。这项努力的结果就是《门罗报告》（Dittrich, Kenneally, and others 2011）。

《贝尔蒙报告》和《门罗报告》共同提供了四条可以指导研究人员进行道德伦理审查的原则：对人的尊重原则、有利化原则、公正原则、对法律和公共利益的尊重原则。在实践中应用这四项原则并不是简单直接的，人们可能需要做出很困难的平衡。尽管如此，这些原则仍有利于权衡利弊，提出研究设计方面的改进建议，并使研究人员能够向彼此和公众解释缘由。

6.4.1 对人的尊重原则

对人的尊重原则在于将人视作独立自主的个体并且尊重他们的愿望。

《贝尔蒙报告》认为，对人的尊重原则由两个不同的部分组成：（1）个体应该被视作独立自主的，（2）缺少独立自主权的个体应该有权获得额外的保护。独立自主，简单来说就是让人们能够自行掌控他们自己的生活。换句话说，对人的尊重原则建议研究人员不应该在未获得同意的情况下采取行动。重要的是，即使研究人员认为发生的事情是无害的，甚至是有益的，也应该遵守对人的尊重原则。对人的尊重原则要求让参与者产生做出决定的想法，而不是由研究人员来做这样的决定。

在实践中，对人的尊重原则被解释为研究人员应该尽可能从参与者那里获得知情同意。知情同意的基本思想是，应以可理解的形式向参与者提供相关信息，然后使其自愿同意参与其中。这些相关信息的每一个术语本身都是大量额外争议和学问的主体（Manson and O’Neill 2007），我将在6.6.1小节中阐述知情同意。

在本章开头的每个事例中，研究人员对参与者都采取了相应的行动，在“3T”项目中使用参与者的数据，在“Encore”项目中使用他

们的计算机对测量任务进行评估，在情绪感染项目中控制他们的行为，这些均没有获得或无视了参与者的知情同意。违反对人的尊重原则不会自动使这些研究在道德伦理方面遭到禁止，对人的尊重原则仅是四项原则之一。但是多考虑一些尊重他人的方式的确可以在道德伦理上改善这些研究。例如，研究人员本可以在研究开始前或结束后从参与者处获得某种形式的知情同意。我在6.6.1小节中讨论知情同意时，会回顾这方面的项目。

6.4.2 有利化原则

有利化原则在于理解和改善研究中所存在的风险/收益状况，然后判断研究是否达到正确的平衡。

《贝尔蒙报告》认为，遵循有利化原则是研究人员对参与者的义务，它涉及两部分：（1）不伤害，（2）最大程度保障有利及最小程度造成伤害（如果无法避免）。《贝尔蒙报告》从古希腊医学家希波克拉底在传统医学伦理中的“不伤害”原则中获得灵感，并且以一种强有力的形式表达出研究人员“不应该伤害一个人，不管这可能会给别人带来什么好处”（Belmont Report 1979）。尽管如此，《贝尔蒙报告》也承认，理解这可能给别人带来什么好处的过程本身也许会让某些人暴露在风险中。因此，不伤害的必要性可能与理解过程的必要性相冲突，导致研究人员偶尔要做出艰难的决定：“在涉及风险的情况下，何时我们可以理所当然地寻求某些利益，何时因风险的存在我们应该放弃相关利益？”

在实践中，有利化原则被解释为研究人员应该实施两个不同的步骤：风险/收益分析，以及随后判定风险和收益是否达到适当的道德伦理平衡。第一个过程主要涉及实质性专业知识的技术问题，而第二个

过程主要是伦理问题，在第二个过程里，实质性专业知识可能不那么有价值，甚至是有害的。

风险/收益分析包含理解与改善研究中的风险与收益。对风险的分析应该包括两个要素：不良事件发生的概率以及这些事件的严重程度。作为风险/收益分析的结果，一名研究人员可以调整研究的设计方案，以降低不良事件发生的概率（比如排除心理脆弱的参与者），或者在其发生后减轻事件的严重程度（比如向有需求的参与者提供咨询服务）。再者，在风险/收益分析的过程中，研究人员需要明确，他们的研究所产生的影响不仅限于参与者，也包括非参与者和社会公众。譬如，请想想雷斯蒂沃与范德里杰特关于奖励对维基百科编辑的影响实验（在第4章讨论过）。在这个实验中，研究人员对一小部分他们认为有价值的编辑给予一些奖励，然后追踪这些人之后对维基百科的贡献，与之对照的是另一部分同样值得奖励的编辑，但研究人员没有给予奖励。试想一下，如果不是仅对一小部分人提供奖励，而是对维基百科的编辑均提供非常多的奖励会是什么样的？尽管这样的设计并不会伤害任何一位参与者，但是它可能破坏整个维基百科的系统奖励机制。换句话说，当进行风险/收益分析时，你不仅应该思考你的工作对参与者的影响，而且应该把整个世界都更多地考虑进来。

接下来，一旦风险被最小化，利益被最大化后，研究人员就应该评估研究是否能够带来一个良好的平衡。伦理学家并不建议单纯地计算成本与收益。特别是，某些风险会致使研究不被允许做实验，无论其收益有多高（比如在历史附录中涉及的塔斯基吉梅毒实验）。与技术性的风险/收益分析不同，第二个步骤需要非常符合道德标准，事实上可以由没有特定专业领域和专业知识的人来实施。事实上，局外人往往相较于局内人更能够注意到各种不同事件，这就是在美国的机构审查委员会中至少需要一位非研究人员参与的原因。在我服务机构审查委员会的经历中，这些局外人能帮助我们防止从众思考。因此，如果你在研究项目中对是否适当进行了风险/收益分析感到疑惑，那么别

去询问你的同事，试着去请教一下非研究人员的意见，他们的答案可能会让你感到意外。

在我们分析的三个事例中，应用有利化原则可能会改善其风险/收益的平衡。譬如，在情绪感染项目中，研究人员本可以尝试排除18周岁以下的用户以及对这项措施反应特别过激的用户。他们本可以通过一些有效的统计学方法尽量减少参与者的人数（具体细节在第4章中已经讨论到）。再者，他们本可以试图监控参与者，并向似乎受到伤害的人提供帮助。在“3T”项目中，研究人员本可以在他们公开数据时附带特别保护措施（尽管他们的程序得到了哈佛机构审查委员会的批准，而这一批准表明在当时这样做是符合常规做法的）；我在6.6.2小节中描述信息化风险时，会提供一些更详细的有关数据公开的建议。最后，在“Encore”项目中，研究人员本可以尽量减少为实现测量目标而创建的风险请求的数量，并且他们本可以排除受到专制政府威胁最大的参与者。每一项可能的改变均需要这些项目的设计做出一些取舍，我的目的不在于建议研究人员做出这些改变，而在于更多地展示有利化原则能够带来的改变。

最后，虽然数字时代通常使风险和收益的平衡更加复杂，但它也使研究人员更容易增加其工作的收益。特别是，数字时代的工具极大地促进了开放和可重复研究，研究人员可以将他们的研究数据和代码提供给其他研究人员，并通过公开发表向公众提供他们的论文结果。开放和可重复研究的这种趋势绝非简单的变化，它为研究人员提供了一种增加研究收益的途径，而不会使参与者面临任何额外的风险（数据共享是一个例外情况，我将在6.6.2小节中详细讨论信息化风险）。

6.4.3 公正原则

公正原则是确保研究的风险和收益能够被公平地分配。

《贝尔蒙报告》认为，公正原则涉及分配研究产生的风险与收益。即在社会环境中不应该仅由某一个小组承担研究成本，而另一个小组获得其产生的全部利益，譬如，19世纪到20世纪初，在医学实验中担任研究对象的负担主要落在穷人身上，而因此获得更好的医疗保健的好处则主要流向了富人。

在实践中，公正原则最初被解释为弱势群体应该被保护，免受研究人员的伤害。换句话说，研究人员不应该被允许故意地侵犯弱势群体。令人不安的是，在过去，大量存在道德伦理问题的研究通常都会涉及弱势群体，包括缺乏教育的和被褫夺公权的公民（Jones 1993）、囚犯（Spitz 2005）以及住院的老弱患者（Arras 2008）。

尽管如此，大约在1990年，对公正的看法开始从保护转向接受（Mastroinanni and Kahn 2001）。譬如，积极分子认为，儿童、妇女和少数民族需要明确纳入临床试验之中，以便这些群体可以从临床试验获得的知识中受益（Epstein 2009）。

除了关于保护与接受的问题之外，公正原则还经常引发对参与者的适当补偿的问题，这也是在医学伦理方面存在激烈争议的问题（Dickert and Grady 2008）。

把公正原则应用到我们的三个事例中，这为我们提供了不同的方式去审视它们。这些研究中没有任何一个向参与者提供了经济补偿。“Encore”项目引起了最为复杂的公正原则问题。尽管有利化原则可能建议排除来自专制政府国家的参与者，但公正原则可能主张让这些人参与进来并从中获益，以此来准确测量互联网审查制度。“3T”项目也引发了争议，因为仅有一组学生承担了研究的负担，而整个社会从中受益。最后，在情绪感染项目中，承担研究负担的参与者是最可能从结果中受益的人口随机样本（即脸谱网用户）。从这个意义讲，情绪感染项目的设计与公正原则是非常一致的。

6.4.4 对法律和公共利益的尊重原则

对法律和公共利益的尊重原则，将有利化原则从具体的研究参与者延伸到了整个利益相关者群体。

第4个也是最后一个指导我们思考的原则就是对法律和公共利益的尊重原则。这项原则来源于《门罗报告》，因此并不为广大社会研究人员所知。《门罗报告》认为，对法律和公共利益的尊重原则隐含在有利化原则之中，但它也认为前者值得被明确考虑。特别是，当有利化原则趋于关注参与者时，对法律和公共利益的尊重原则明确鼓励研究人员应该有更为广阔的想法和法律方面的考量。

在《门罗报告》中，对法律和公共利益的尊重原则包含两个明确的部分：（1）合规性，（2）基于透明的问责制。合规性意味着研究人员应该尝试识别并遵守相关法律、合同和服务条款。例如，合规性意味着，想要获取网站内容的研究人员应该阅读并考虑该网站的服务条款协议。尽管如此，也有可能存在违反服务条款的情况；请记住，对法律和公共利益的尊重原则仅是四项基本原则之一。譬如，威瑞森电信（Verizon）与AT&T（美国电话电报公司）曾一度有服务条款阻止客户对其进行批评（Vaccaro et al. 2015）。我认为研究人员应该遵从服务条款。在理想情况下，如果研究人员违背了服务条款协议，那么他们应该公开解释这样做的理由（参见Soeller et al. 2016），正如基于透明的问责制所建议的那样。但是这样的公开化可能将研究人员暴露在附加的法律风险下，譬如，美国《计算机欺诈和滥用法》（CFAA）可能将违反服务条款定义为非法（Sandvig and Karahalios 2016; Krafft, Macy, and Pentland 2016）。这个简短的讨论表明，伦理审议中的合规性可能引发更为复杂的问题。

除了合规性以外，对法律和公共利益的尊重原则也鼓励基于透明的问责制，这意味着研究人员应该对各阶段的研究目标、方法以及结

果都很明确，并且对其行为负责。从另一个角度来想，基于透明的问责制试图禁止研究团体的秘密行为。这种基于透明的问责制使公众在道德伦理争议中发挥了更广泛的作用，这对伦理和实践两方面都很重要。

将对法律和公共利益的尊重原则应用于这里所考虑的三项研究中，可以说明研究人员在涉及法律时所面临的问题的复杂性。譬如，格林默曼指出情绪感染项目在马里兰州可能是非法的（Grimmelmann 2015）。特别是在2002年马里兰州议会通过了917法案（Maryland House Bill 917），将《通则》保护拓展到在马里兰州进行的所有研究，并且这种保护与研究的资金来源无关（很多专家认为情绪感染项目的主体并不是属于联邦法律《通则》下的主体，因为该项目由脸谱网发起，而该机构并不受到美国政府的研究资助）。尽管如此，一些学者认为马里兰州917法案本身也属违宪（Grimmelmann 2015, pp.237-238）。社会研究人员并非法官，因此他们也不必理解或评估全美国50个联邦州的法律是否符合宪法。在国际项目中，这样的复杂性变得更加严重。譬如“Encore”项目涉及170个国家的参与者，要从合规性的角度考虑是异常困难的。为了回应模糊的合法环境，研究人员可能通过第三方监管其工作来获益，以防研究中的无意违法，第三方既是法律要求的建议来源，也是对个人的保护。

另一方面，这三项研究均将其结果发表于学术期刊，以实现基于透明的问责制。事实上，情绪感染项目的结果以公开的形式发表，所以研究机构或者社会大众能够获知其设计方案和研究结果。一个简单而粗略的用来评估基于透明的问责制的方法是询问你自己：当我的研究过程在我家乡的报纸头版上被提及时，我是否感到心安？如果答案是否定的，那么就说明你的研究设计可能需要做出调整。

总而言之，《贝尔蒙报告》和《门罗报告》提出了四项可用于评估科学研究的原则：对人的尊重原则、有利化原则、公正原则以及对

法律和公共利益的尊重原则。应用这四项原则在实践中并不是简单直白的，它可能需要更为复杂的权衡。譬如，在决定是否向情绪感染项目的参与者汇报该项目的情况时，对人的尊重原则就会鼓励研究人员告知，但是有利化原则就不会鼓励这样做（如果汇报本身可能造成伤害的话）。并不存在一种自动的方式可以权衡相互冲突的原则，但是这四项原则帮助我们理解了如何做出取舍，对研究设计方案的调整给出了建议，还能让研究人员对他人和社会公众解释其缘由。

6.5 两种道德框架

大多数关于实验研究的道德伦理的争议都会减少结果主义与义务论之间的分歧。

对人的尊重原则、有利化原则、公正原则以及对法律和公共利益的尊重原则这四项原则本身来源于更为抽象的两种道德框架：结果主义与义务论。理解这两种框架有所裨益，因为它们将可以让你识别并推断出伦理研究中最根本的紧张关系之一：使用潜在的不道德手段达到道德目的。

结果主义来源于杰里米·边沁与约翰·穆勒（John Stuart Mill）的思想，关注于采取相关行动使世界上的国家变得更好（Sinnott-Armstrong 2014）。有利化原则旨在权衡风险和利益，是结果主义者更为深层次的思考来源。义务论来源于康德的思想，关注于道德义务，它与其所收到的结果无关（Alexander and Moore 2015）。对人的尊重原则重视参与者的自主权利，它是一种道德义务的深层思考来源。一个简单而粗略的用以区分这两种框架的方法是：道德义务论者关心过程，而结果主义者关心结果。

要了解这两种框架如何运作，可以参考知情同意。这两个框架均能用以支持知情同意，但是有不同的理由。结果主义者的论证观点在于，对于禁止那些不能准确权衡风险与预期利益的研究，知情同意能保护参与者免受伤害。换言之，结果主义者认为支持知情同意能帮助参与者免受不良结果的影响。尽管如此，对义务论者来说，其论证观点在于，知情同意关注研究人员有义务尊重参与者的自主决定。基于

这些方法，纯粹的结果主义者可能愿意在没有风险的情况下放弃获得知情同意的要求，而纯粹的道德义务论者可能不会。

结果主义和义务论都提供了重要的道德洞察力，但每个都可以被视为荒谬的极端。对于结果主义，其中一个极端例子是移植。试想一个医生有5名因器官衰竭而濒临死亡的病患，而另外一个人的器官可以拯救这5个人。在确定的条件下，一名结果主义的医生会允许甚至要求杀死那位健康的人去获取他的器官。他完全只看结果，并不考虑过程，这是有缺陷的。

同理，义务论也一样拥有荒唐的极端，例如一个关于定时炸弹的例子。试想一个警察抓捕到一名恐怖分子，恐怖分子知道一枚能危及数百万人生命的定时炸弹的位置。一个信奉义务论的警察可能不会通过欺骗来从恐怖分子口中获知定时炸弹的位置。他完全只看过程，并不考虑结果，也一样是有缺陷的。

在实践中，大多数社会研究人员含蓄地融合了这两种道德框架。我们注意到，这两种道德框架的融合可以帮助我们理解为什么存在如此多的伦理争议，因为某些人更趋近于结果主义者而另一些人更趋近于义务论者，且双方无法取得更多的进展。结果主义者往往提出有关结果的论点，而这样的论点并不被义务论者所接受，他们更忧心于过程。同理，义务论者倾向于提供有关过程的论点，而这也不被结果主义者所接受，他们更关注结果。结果主义者与义务论者就此渐行渐远。

这种争议的一个解决方案是让社会研究人员发展出一种一致的、道德坚实的、易于操作的结果主义与义务论的融合体。很遗憾，这不太可能发生，哲学家已经被这个问题困扰了很长一段时间。尽管如此，研究人员仍旧能够使用这两种框架以及四项原则，为道德伦理挑战提供理由，明确利弊权衡以及改善研究设计方案。

6.6 困难面

四项原则，即对人的尊重原则、有利化原则、公正原则以及对法律和公共利益的尊重原则；两种道德框架，即结果主义与义务论，这些应该能帮助你厘清任何研究中所面临的道德伦理问题。尽管如此，基于在本章之前描述的数字时代研究特性以及我们迄今已考虑到的伦理争议，我认为存在四个特定的困难面：知情同意、理解与管理信息化风险、隐私权以及面对不确定性所做出的决策。在接下来的几个小节中，我将描述这四项要点的具体内容并提供如何处理它们的一些建议。

6.6.1 知情同意

研究人员应该、可以并且需要遵循如下规则：在大多数研究中获得某种形式的同意。

知情同意是研究道德伦理方面的一个基本想法，有些人可能说这是一种近乎强迫的想法（Emanuiel, Wendler, and Grady 2000; Manson and O’Neill 2007）。最简单的伦理研究要求：所有事情都需要知情同意。尽管如此，这个简单的规则对于既有的道德原则、道德规则或是研究方法都不一致。取而代之，研究人员应该、可以并且需要遵循更为复杂的规则：在大多数研究中获得某种形式的同意。

首先，为了跳出关于知情同意的过分简单化的想法，我想告诉你更多关于研究歧视的实地调查。在这些研究中，虚假申请人具有不同的特征，比如一些男性和一些女性申请不同的工作。如果其中一类申

请者更频繁地受雇，那么研究人员可以归纳认为雇佣过程中可能存在歧视。对于本章的目的来说，实验最重要的事情在于参与者，即雇主，在实验的过程中从没有同意参与实验。事实上，这些参与者都被积极地欺骗了。然而，在17个国家进行的这种研究歧视的实地调查有117项之多（Riach and Rich 2002; Rich 2014）。

采用实地调查研究歧视的研究人员已经确定了这些研究的四项特征，总体而言，使其符合道德标准：（1）对雇主有限的伤害；（2）拥有可靠的歧视衡量措施，进而可获取巨大的社会效益；（3）衡量歧视的其他方法有其自身弱点；（4）欺骗行为并没有严重违反规定的形式（Riach and Rich 2004）。其中每一项条件都是关键的，一旦其中一条不被满足，那么这个伦理事例将面临更多挑战。其中三项特征来源于《贝尔蒙报告》涉及的道德原则：有限伤害（对人的尊重原则、有利化原则），获取最大利益，相较而言其他方法有其不足（有利化原则、公正原则）。最后一项特征，遵守相关规则，则来源于《门罗报告》中提及的对法律和公共利益的尊重原则。换言之，就业申请其本身是一个已存在一些可预期欺骗的环境。因此，这些实验并不会侵犯既有的原始道德观。

除了这种基于原则的论点之外，多数机构审查委员会认为，这些研究虽然缺乏知情同意，但与现有规则还是相一致的，特别是与《通则》第46章116条*d*部分中所描述的相一致。最后，美国法院也支持在实地调查中缺乏同意和使用欺骗来衡量歧视的行为（No. 81-3029. United States Court of Appeals, Seventh Circuit）。因此，在未经同意的情况下进行实地调查符合既有的道德原则与规则（至少符合在美国的规则）。这个理由被广大社会研究团体、多数机构审查委员会以及美国上诉法院所支持。所以，我们必须放弃“所有事情都需要知情同意”这一简单规则。这不是研究人员必须遵循的规则，也不是他们应该遵循的规则。

跳出“所有事情都需要知情同意”这一框架，研究人员面临着一个棘手的问题：对于不同种类的研究，究竟需要何种形式的同意？当然，围绕这个问题一直存在很大的争议，尽管其中大部分讨论都是在模拟时代的医学研究背景下进行的。尼尔·埃亚勒（Nir Eyal）在2012年将这些争论总结为：

干预的风险越大，越是能影响或定义“关键的生活选择”；干预的价值载重和争议性越大，干预直接影响的身体领域就拥有越多隐私；研究人员越是面临冲突与无监管状态，对强有力的知情同意的需求就越高。在其他情况下，对非常有力的知情同意的需求，对所有事情都要知情同意的需求，要少得多。在这种情况下，高成本可能很轻松地超过了其本身的需求。

这次争论得出的一个重要见解是，知情同意并非全部，也不是一无是处，有更强形式或者更弱形式的同意存在。在一些场景中，强有力的知情同意似乎非常必要，但在另一些情况下，弱一些的知情同意的形式可能更合适。接下来，我会描述三种研究人员可能努力去获取知情同意的理由，并且给这些事例提供一些选项。

首先，有时向参与者提出获取知情同意可能增加其面对的风险。譬如，在“Encore”项目中，寻求生活在专制政府下的人们的知情同意，用他们的计算机测量互联网审查制度，可能会让他们身处风险更高的境地。当他们的同意导致风险增加时，研究人员可以确保他们正在做的事情的信息是公开的，参与者可以选择退出。同样，研究人员也可以寻求代表参与者的组织机构（例如非政府组织）。

第二，有时在研究开始之前获得完全知情同意可能会损害研究的科学价值。譬如，在情绪感染项目中，如果参与者获知研究人员正在做一个关于情绪的实验，就可能改变他们的行为。阻碍参与者获取信息，甚至欺骗他们，在社会研究中并不少见，特别是在心理学的实验研究中。如果在研究开始之前无法获取知情同意，那么研究人员可以

（通常也这样）在研究结束后向参与者进行事后说明。这样的事后说明通常包括阐述实验的过程、对任何伤害实施补救，以及事后获取同意。尽管如此，当事后说明本身可能对参与者造成伤害时，有关是否在实地调查中对参与者进行说明，往往存在一些争议（Finn and Jakobsson 2007）。

第三，有时向受影响的每个人争取知情同意在逻辑上是不切实际的。譬如，试想一下，如果你是一位期望研究比特币区块链技术（比特币是一种虚拟货币，区块链是比特币交易的公开记录）的研究人员。不幸的是，获取每一个使用比特币的人的知情同意是不可能的，因为大多数人都是匿名的。在这样的情况下，研究人员可以尝试联系一个比特币使用者作为样本，向其获取知情同意。

这三个研究人员可能无法取得知情同意的原因，即增加风险、损害研究目标以及逻辑限制，都不是研究人员努力争取获得知情同意的全部原因。我所建议的解决方案——向公众宣传研究成果、允许退出的选择、寻求第三方的同意、事后说明，以及征求参与者样本的同意，在所有情况下可能都无法实现。此外，即使这些替代方案是可行的，它们可能不足以用于既定的研究。尽管如此，这些例子所展示的知情同意既不是全部，当然也不会是一无是处。创造性的解决方案可以在无法获取所有受影响方完全知情同意的情况下，改善其道德平衡。

总而言之，比起“所有事情都需要知情同意”，研究人员应该、可以并且也需要遵循一个更复杂的准则：在大多数研究中获得某种形式的同意。就原则而言，出于对人的尊重原则，知情同意既不是必要的，也不是充分的（Humphreys 2015, p.102）。当我们考虑研究伦理时，对人的尊重原则仅是需要权衡的一项原则；它不应该自动凌驾于有利化原则、公正原则以及对法律和公共利益的尊重原则之上，在过去的40年中，伦理学家反复指出这一观点（Gillon 2015, pp.112-

113）。就道德框架而言，对每一件事都获得知情同意就是完全站在义务论的角度去考虑问题，正如定时炸弹那个例子中的警察一样（参见6.5节）。

最终，作为一个实际问题，如果你正在考虑在没有任何同意的情况下进行研究，那么你应该知道自己正处于灰色地带。请注意回顾研究人员为了在未经同意的情况下进行歧视实验研究而提出的道德论点。你的理由足够强吗？因为知情同意是许多非专业道德理论的核心，你应该知道，你很可能会被要求为你的决定进行辩护。

6.6.2 理解与管理信息化风险

信息化风险是社会研究中最普遍的风险，它急剧增加，也是最难理解的风险。

在数字时代的社会研究中，第二个道德挑战就是信息化风险，一种因泄露某些信息而带来的潜在伤害（National Research Council 2014）。来自个人信息的泄露所产生的信息化伤害可以体现在经济方面（比如丢失工作）、社会方面（比如身处尴尬境地）、心理方面（比如抑郁），甚至是犯罪方面（比如因非法行为被捕）。不幸的是，在数字时代，这种信息化伤害急剧增加，因为我们的行为中藏有更多的信息。相较于模拟时代下社会研究所关心的风险，比如自然风险，信息化风险更难被理解与管控。

研究人员减少信息化风险的一项措施是数据“匿名化”。“匿名化”是从数据中移除诸如姓名、住址以及电话号码等显著个人信息的过程。尽管如此，该方法远不如许多人以为的那样有效，事实上，这种方法的深度和广度均受到限制。因此，无论何时，在描述“匿名

化”时，我都将使用双引号来提醒你这样的过程只是一种表面的匿名，并非真正的匿名。

一个关于“匿名化”失败的生动的例子来自20世纪90年代晚期的马萨诸塞州（Sweeney 2002）。团体保险委员会（GIC）是一个政府机构，对缴纳健康保险的所有联邦雇员负责。通过这样的工作，团体保险委员会搜集到了有关联邦雇员的大量详尽的健康记录。为了促进研究，团体保险委员会决定将这些记录发布给研究人员。尽管如此，他们并不会公布所有的数据；相反，他们通过删除姓名和住址等信息来“匿名化”这些数据。但是，他们留下了自己认为可能对研究人员有用的其他信息，例如人口统计信息（邮编、出生日期、族裔以及性别），医疗信息（医生访问时间、诊断以及实施步骤）（图6.4）（Ohm 2010）。不幸的是，这种“匿名化”并没有充分保护这些数据。

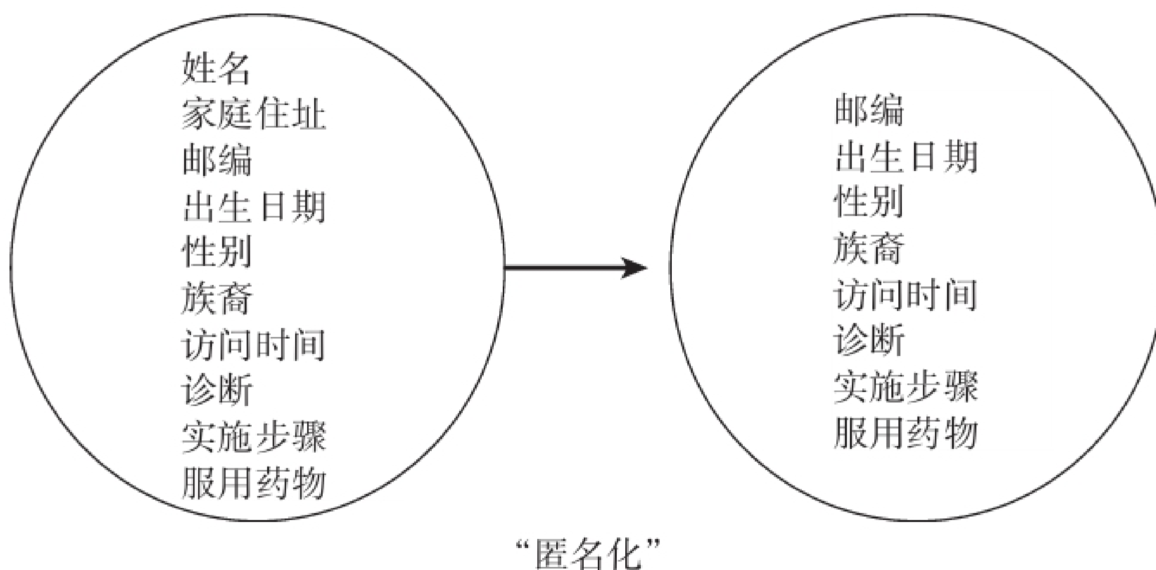


图6.4 “匿名化”是移除显著认证信息的过程。比如，当公开联邦雇员的医疗保险记录时，马萨诸塞州团体保险委员会从文件中移除姓名以及家庭住址等信息。

为了说明团体保险委员会“匿名化”的缺点，拉塔尼娅·斯威尼（Latanya Sweeney），一名麻省理工学院的研究生，支付了20美元购

买了马萨诸塞州州长威廉·韦尔德（William Weld）的故乡剑桥市的投票记录。这些投票记录包括了诸如姓名、住址、邮编、生日以及性别等信息。事实上，医疗数据文件和投票记录有许多共同的信息，包括邮编、生日和性别，这意味着斯威尼也可以将其联系起来。斯威尼知道韦尔德的生日是1945年7月31日，根据投票记录，剑桥市只有6个人有同样的生日。再进一步，这6人中只有3人是男性。接下来，3位男性中只有一人跟韦尔德的邮编一致。因此，根据投票数据显示的内容，任何人都能够将韦尔德的生日、性别和邮编信息与医疗记录联系起来找出韦尔德。本质上，这三条信息在数据中像是他的一个独特的指纹信息。通过这样的结果，斯威尼能够定位出韦尔德的医疗记录，为了告知韦尔德她的成就，斯威尼向他寄送了一份数据拷贝（Ohm 2010）。

斯威尼的工作说明了“再识别攻击”的基本结构，这是一个计算机安全领域的术语。在这些攻击中，两个数据库本身都没有显示敏感信息，但两个数据库是相互关联的，通过这种联系，攻击者使得敏感信息被暴露。

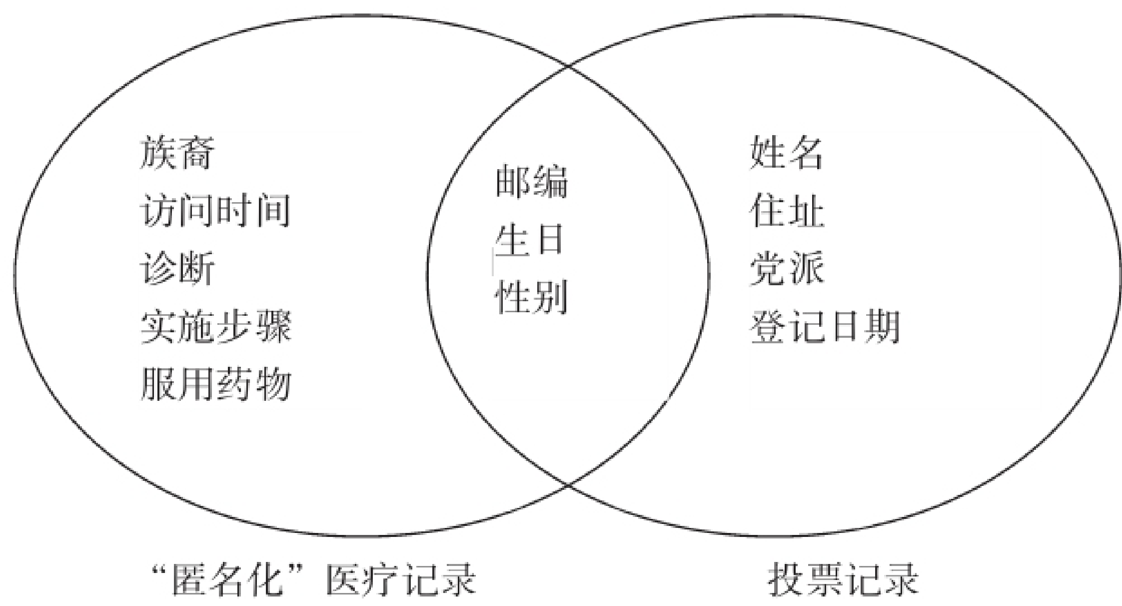


图6.5 “匿名化”数据的再识别。拉塔尼娅·斯威尼通过结合投票记录和“匿名化”医疗记录，寻找威廉·韦尔德州长的医疗记录。来源于Sweeney（2002），图

1。

为了回应斯威尼的工作和其他相关工作，研究人员现在通常在整个“匿名化”过程中删除更多的信息，即所谓的“个人识别信息”（PII）（Narayanan and Shmatikov 2010）。此外，许多研究人员现在认识到，即使在“匿名化”之后，某些数据（如医疗记录、财务记录、有关非法行为的调查问卷回答）也可能过于敏感。我要讲的例子表明社会研究人员需要改变他们的想法。作为第一步，假设所有数据都有可能被识别，并且所有数据都可能是敏感的，这是明智的。换句话说，与其认为信息化风险适用于一小部分项目，我们还不如假设它在某种程度上适用于所有项目。

网飞奖表明了这种重新定位的两个方面。正如第5章所描述的，网飞公司公布了由近50万名会员提供的1亿条电影评级信息，并且公开征集来自世界各地的人提交的算法，以提高其推荐电影的能力。在公布这些数据之前，网飞公司移除了所有显著的个人认证信息，比如姓名等。他们还增加了一项特别措施，在一些记录中引入了轻微的干扰项（比如将某些评级由4星改为3星）。尽管如此，网飞很快发现，虽然他们付出了相应的努力，但数据仍然不是匿名的。

仅在他们公开这些数据的两周后，阿尔温德·纳拉亚南与维塔利·施马季科夫就表明，了解特定人群的电影喜好是可能的。其窍门在于采用与斯威尼手法相似的再识别攻击：把两个信息源合为一体，一个信息源具有潜在的敏感信息，但没有明显的识别信息，另一个包含人的身份信息。虽然各自的信息源可能都是相对安全的，但是它们被合并后，就增加了信息化风险。在网飞数据的例子中，就发生了这样的事。试想一下，当我选择与同事分享我关于动作电影与喜剧电影的想法时，我并不会愿意分享我对宗教和政治类电影的看法。我的同事可能利用我所分享的看法去网飞数据库中找到相关信息。我所分享的信息可能像威廉·韦尔德的生日、邮编、性别信息那样，成为一个独特的指纹。他们可以了解到我对所有电影的评价，包括我选择不分享

的电影。除了这种针对单人的目标攻击之外，纳拉亚南与施马季科夫还表明，通过将网飞数据与某些人选择在互联网电影数据库（IMDb）上发布的电影评级数据合并，可以进行广泛攻击，即涉及多人的攻击。很简单，任何特定人员的独特指纹信息，即使是他们的电影评级信息，都可以用于识别他们。

尽管网飞数据可以被用于再识别攻击或广泛攻击，但是它仅可能是低风险的。毕竟，电影评级信息看起来并不那么敏感。虽然这在通常情况下是正确的，但对集中了50万人的数据库来说，电影评级信息可能就相当敏感。事实上，作为对再识别的回应，一名未公开身份的女同性恋者加入了对网飞公司的集体诉讼中。

网飞奖数据的再识别说明，所有数据都有可能被识别，并且所有数据都可能是敏感的。此时，你可能会认为这只适用于那些与人有关的数据。令人惊讶的是，其实不是这样的。为了回应《信息自由法》的请求，纽约市政府公开了2013年纽约市所有的出租车行驶记录，包括其接客与落客的时间、位置以及付款金额等信息（回顾第2章，法伯使用了类似的数据来检验劳动经济学中的重要理论）。这些有关出租车行程的数据可能看起来没什么问题，因为它们并不涉及人的信息，但安东尼·托卡（Anthony Tockar）意识到这些出租车数据库实际上包含了许多有关人的潜在敏感信息。为了确认这一点，他浏览了午夜到早上6点从纽约一家大型脱衣舞酒吧出发的行程记录，并找到其落客位置。这项搜索实质上找出了一些经常光顾这个酒吧的人的住址（Tockar 2014）。很难想象市政府在公布数据时是否考虑到了这一点。事实上，用同样的办法可以找到去这座城市任何地方，包括去诊所、政府大楼或者是宗教场所的人的家庭住址。

网飞奖和纽约市出租车数据的这两种情况表明，相对有经验的人也可能无法正确地估计他们发布的数据中存在的信息化风险，而这些案例并非个例（Barbaro and Zeller 2006; Zimmer 2010;

Narayanan, Huey, and Felten 2016）。而且，在许多这样的情况下，有问题的数据仍然可以被在线免费获取，这表明撤销已发布的数据是非常困难的。总的来说，这些例子以及计算机科学关于隐私的研究引出了一个重要的结论。研究人员应该假定所有数据都有可能被识别，并且所有数据都可能是敏感的。

不幸的是，并不存在简单的解决方案，也就是说所有的数据都可能被识别，并且所有数据都可能是敏感的。尽管如此，在你的数据工作中，其中一项能够降低信息化风险的措施是创建并遵循一个数据保护计划。该项计划能够降低你泄露数据的概率，并且在数据泄露发生后能够降低伤害。随着时间的推移，数据保护计划的具体项目，包括能够使用的加密形式，都在改变。英国数据服务中心（UK Data Services）有效地归纳了数据保护计划所必备的5项要素，他们称之为“5个安全”：项目安全、对象安全、数据安全、设置安全、成果安全（表6.2）（Desai, Ritchie, and Welpton 2016）。这5项安全措施中的任何单独一项都不能提供完美的保护措施。但是将它们组合在一起，则可以有力降低信息化风险。

表6.2 “5个安全”是设计和执行数据保护计划的原则

安全措施	行动
项目安全	对涉及道德伦理的项目数据采取限制措施
对象安全	访问仅限于可信任的数据人员（例如，经过道德培训的人）
数据安全	尽可能将数据去标识并汇总
设置安全	对储存于计算机中的数据采取适当的物理（例如，锁闭的房间）和软件（例如，密码保护、加密）保护
成果安全	审查研究成果以防止意外隐私泄露

除了在使用数据时保护数据以外，研究过程中信息化风险特别突出的一个步骤是与其他研究人员共享数据。科学家之间的数据共享是

科学探索的核心价值，并且它能够在很大程度上促进知识的进步。以下是英国下议院对数据共享重要性的看法（Molloy 2011）：

如果研究人员要重现、验证文献中发表的结果并在其基础上建立其他研究，获取数据是至关重要的。因此我们必须假定，除非拥有很强的其他理由，否则数据应该被充分披露并公开。

至此，当与其他研究人员共享数据时，你可能增加了你的信息化风险。因此，似乎在与其他科学家分享数据的义务和为参与者减少信息化风险的义务之间，共享数据这一行为带来了基础性的紧张关系。幸运的是，这种矛盾并不像看起来那么严重。相反，最好将数据共享视为一个连续统一体，这个连续统一体的每一点提供了不同的社会收益与参与者风险的组合（图6.6）。

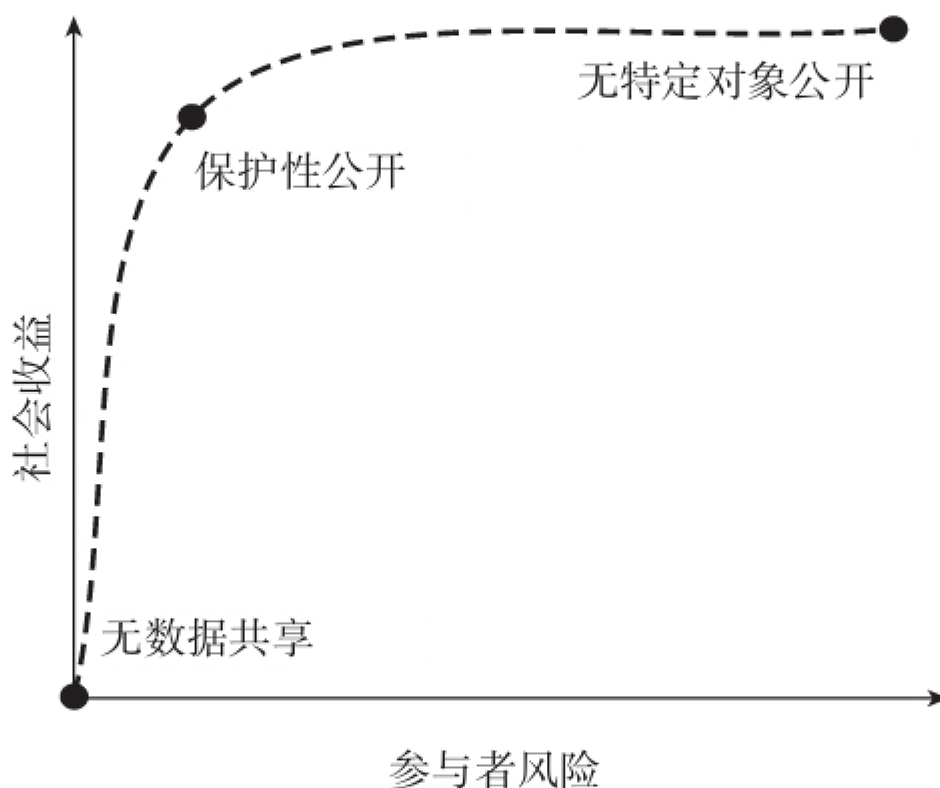


图6.6 数据共享的策略位于一个连续统一体之中。你应该在这个连续统一体中的哪个部分，取决于你的数据具体细节，第三方监管可能会帮助你决定案例中风险收益

的适当平衡。这条曲线的确切形状取决于数据和研究目标的具体情况（Goroff 2015）。

在一种极端情况下，你可以不对任何人共享你的数据，这样的话参与者风险也就最小化了，相应社会收益也最小。在另一种极端情况下，你可以无特定对象公开，数据被“匿名化”并被所有人使用。相较于不公开数据，这种无特定对象公开能够提供更高的社会收益，但也随之给参与者带来了更高的风险。在这两种极端情况之间的混合范围里，存在一种我称之为保护性公开的方法。在这样的方法下，你可以将数据与符合特定标准并同意受某些规则约束的人共享（比如在机构审查委员会和数据保护计划的监管下）。这种保护性公开的方法提供了许多公开的好处，并减少了无特定对象公开的风险。当然，这样的方法也会产生很多问题，比如，谁能获得这样的权限，在什么样的条件下，能使用多久，谁又为这样的保护性公开所带来的监管成本埋单等，但这些都是可以被解决的。事实上，在有些地方，现在已经有相应的保护性公开方法被研究人员采用了，比如密歇根大学的校际政治及社会研究联盟（ICPSR）的数据档案。

那么，你在研究中将采取何种数据处理措施？非公开，保护性公开，还是无特定对象公开？这个取决于你的数据的具体情况，研究人员必须权衡四项原则。从这个角度看，数据共享并不是一个独特的道德难题，它只是研究人员必须找到合适的道德平衡的许多方面之一。

一些批评家普遍反对数据共享，在我看来，他们专注于风险，这无疑是对的，但他们忽略了它的好处。因此，为了鼓励关注风险与收益，我想提供一个类比。每年，因为汽车所产生的事故都会造成数以千计的人死亡，但是我们不会试图禁止驾车。事实上，禁止驾车的呼吁是荒谬的，因为驾车也能为我们带来许多美好的事情（不能因噎废食）。相反，社会可以限制谁能驾车（例如，需要达到某个年龄并通过某些测试），以及他们如何驾车（例如在限速的情况下），社会也有负责执行这些规定的人员（例如警察），我们会惩罚那些违反规则

的人。同样，社会管理驾车问题的平衡思想也可以用于数据共享的过程。也就是说，我认为我们可以通过降低风险和提高数据共享收益，获取最大的进展，而不是为数据共享提供绝对的支持或反对论据。

总而言之，信息化风险增长迅速，并且很难预测和衡量。因此，最好假设所有的数据都有可能被识别，并且所有的数据都可能是敏感的。为了在研究过程中减少信息化风险，研究人员可以创建并遵循一些数据保护措施。另外，信息化风险不会阻止研究人员从其他科学家处获得共享数据。

6.6.3 隐私权

隐私权是让信息适当流通的权利。

第三个研究人员应该努力完善的方向是隐私权。劳伦斯非常简明地指出：“隐私权也应该像人一样受到尊重。”（Lowrance 2012）尽管如此，隐私权是一个众所周知的混乱的概念（Nissenbaum 2010）。因此，在尝试做出有关研究的特定决策时，使用它很困难。

考虑隐私权的常见方式是公/私二分法。通过这样的思考方法，如果信息可以公开获取，那么研究人员使用它就不用在意侵犯了公民的隐私权。然而使用这样的方法会产生问题。例如，在2007年11月，科斯塔斯・帕纳戈普洛斯（Costas Panagopoulos）向三个城镇的每个人发出了关于即将举行选举的信件。在艾奥瓦州的蒙蒂塞洛和密歇根州的霍兰这两个小镇，帕纳戈普洛斯在报纸上保证/威胁将要公布一份参与投票的人的名单。在另一个小镇，艾奥瓦州的伊利，帕纳戈普洛斯保证/威胁公布一份没有投票的人的名单。这些措施旨在引发自豪感与羞耻感（Panagopoulos 2010），因为这些情绪在早期研究中被发现会影响投票率（Gerber, Green, and Larimer 2008）。在美国有关谁参

与投票、谁未参与投票的信息是公开的，并且任何人都可以获取。因此，有人可能会争辩说，因为这个投票信息已经公开了，所以研究人员在报纸上公布它没有问题。但这个论点对某些人来说则会引起不适。

正如这个例子所说明的那样，公/私二分法太过愚钝了（Boyd and Crawford 2012; Markham and Buchanan 2012）。思考隐私权的一种更好的方式是情境完整性，这是一种专门用于处理数字时代问题的方法（Nissenbaum 2010）。情境完整性关注的是信息流通，而不是将信息视为公共或私人信息。引述尼森鲍姆（Nissenbaum）的话："隐私权既不是保密权，也不是控制权，而是适当流通个人信息的权利。"

情境完整性的基本概念是与情境相关的信息化规范（Nissenbaum 2010）。这就是管理特定情境中信息流通的规范，它们由三个要素决定：

- 角色（主体、发送者、接收者）；
- 属性（信息类型）；
- 传输原则（信息流通的限制）。

因此，作为一名研究人员，当你正在决定是否未经允许使用数据时，它能够帮助你询问自己，这样做是否侵犯了与情境相关的信息化规范。回到帕纳戈普洛斯的例子来，在这个例子中，外部研究人员在报纸上公开选民或非选民名单，这似乎可能会违背信息化规范。这可能不是人们期望的信息流通方式。事实上，帕纳戈普洛斯没有执行他的保证/威胁，因为当地选举官员从这些信件中追查到他，并劝他说这并不是一个好的想法（Issenberg 2012, p. 307）。

与情境相关的信息化规范，也可以帮助评估我在本章开始时讨论的情况，即2014年在西非埃博拉疫情暴发期间使用手机通话记录追踪人口流动性的例子（Wesolowski et al. 2014）。在这样的环境设定中，我可以想到两种不同的情境：

- **情境1：** 发送完整的通话记录数据（属性）；给不完全合法的政府（角色）；用以应对未来任何可能的使用（传输原则）。

- **情境2：** 发送部分匿名记录（属性）；给受尊敬的大学研究人员（角色）；用以应对埃博拉疫情，并接受大学机构委员会的监督（传输原则）。

尽管在这两种情境下，通话数据都从移动通信公司流出，但鉴于角色、属性和传输原则之间的差异，这两种情况的信息化规范并不相同。只关注其中一个参数可能导致过于简单的决策。事实上，尼森鲍姆强调，这三个参数不能缩减为其他两个参数，任何单一参数也不能单独定义信息化规范。信息化规范的三维性质解释了为什么过去的努力，即侧重于属性或传输原则的努力，在保护隐私方面效果不佳。

用与情境相关的信息化规范来指导决策的一个挑战是，研究人员可能不会提前知道它们，并且它们很难衡量（Acquisti, Brandimarte, and Loewenstein 2015）。进一步来说，即使一些研究人员会违反与情境相关的信息化规范，也并不自动意味着其研究不应该进行。事实上，尼森鲍姆著作的第8章完全阐明了“为了好事而破坏规则”。尽管存在这些复杂性，但是与情境相关的信息化规范仍然是推断隐私权相关问题的有用方式。

最后，隐私权的问题是我所见过的优先考虑对人的尊重原则与优先考虑有利化原则的研究人员之间普遍存在误解的地方。试想一下，公共卫生研究人员为了防止新型传染病的传播，秘密观察了正在洗澡的人们。关注有利化原则的研究人员在乎这项研究对社会带来的好

处，并且可能会争辩说，如果研究人员在没有被发现的情况下进行了偷看行为，那么参与者就没有受到伤害。另一方面，优先考虑对人的尊重原则的研究人员关注的是研究人员没有对人的起码尊重，并可能认为这侵犯了参与者的隐私权，造成了伤害，即使在参与者没有意识到他们被偷看的情况下。换句话说，对这些研究人员来说，侵犯人们的隐私权本身就是一种伤害。

总而言之，在考虑隐私权问题时，这有助于我们超越过于简单化的公/私二分法，而采用与情境相关的信息化规范，其由三个要素组成：角色（主体、发送者、接收者），属性（信息类型），以及传输原则（信息流通的限制）（Nissenbaum 2010）。一些研究人员根据隐私侵害可能导致的伤害来评估隐私权，而其他研究人员则认为侵犯隐私权本身就是一种伤害。许多数字系统的隐私概念随时间而变化，因人而异，并且因情况而异（Acquisti, Brandimarte, and Loewenstein 2015），因此隐私权很可能在未来某些时候成为研究人员在伦理决策中遭遇困难的根源。

6.6.4 面对不确定性做出决策

不确定性不一定导致无所作为。

我期望研究人员努力的第四个也是最后一个领域是面对不确定性做出决策，也就是说，在哲学化和权衡利弊、研究伦理问题后，决定做什么和不做什么。不幸的是，这些决策往往基于不完整的信息。譬如，当设计“Encore”项目时，研究人员可能希望知道它会导致某人被警方找上门的可能性。或者在设计情绪传染项目时，研究人员可能希望知道它引发某些参与者抑郁的可能性。这些概率可能非常低，但在研究发生之前它们是未知的。而且，因为这两个项目都没有公开追踪有关不良事件的信息，所以其概率仍然不为众人所知。

在数字时代，不确定性并不是社会研究所特有的。当《贝尔蒙报告》描述了风险与收益的系统评估时，它明确承认这些很难精确量化。尽管如此，这些不确定性在数字时代更为严重，部分原因是我们对这类研究的经验较少，另外一部分原因在于其研究本身的特点。

鉴于这些不确定因素，有些人似乎主张“安全性高于遗憾的产生”，这是预防原则的口语化解释。虽然这种方法看似是合理的，甚至可能是明智的，但它实际上可能会造成伤害。它让研究环境变得冷淡，并且导致人们对局势的看法过于狭隘（Sunstein 2005）。为了更好地理解预防原则的问题，让我们回到情绪感染的例子中。实验计划涉及大约70万人，实验中肯定有人会受到伤害。但是，这个实验也有可能产生对脸谱网用户和社会有益的知识。因此，虽然允许实验有风险（正如已经充分讨论过的那样），但阻碍实验也有风险，因为实验可能会产生宝贵的知识。当然，选择做还是不做实验并不是在实验发生时进行的。对实验设计有很多修改方式，它们可能将其带入不同的道德平衡。然而，在某些时候，研究人员可以在做研究和不做研究之间做出选择。行动和不行动都有风险，仅仅关注行动的风险是不恰当的。原因很简单，并不存在完全无风险的方法。

跳出预防原则的限制，在面对不确定性时做出决定的一个重要方法是最小化风险标准。该标准试图将特定研究的风险和参与者在日常生活中承担的风险（例如运动或驾车）进行比较（Wendler et al. 2005）。这种方法是有价值的，因为评估是否符合最低风险标准比评估实际风险水平更容易。譬如，在情绪感染项目中，在研究开始之前，研究人员可以将实验中新消息反馈的情绪内容与脸谱网上的其他新消息反馈的情绪内容进行比较。如果它们是相似的，那么研究人员就可以在符合最小风险标准的情况下进行实验（Meyer 2015）。即使不知道风险的绝对程度，他们也可以做出这个决定。该方法同样可以应用于“Encore”项目中。最初，“Encore”项目触发了对已知敏感网站的请求，例如那些专制政府所禁止的政治党派网站。因此，这对

某些国家的参与者来说风险不小。正因如此，“Encore”的修订版本仅向推特、脸谱网和优兔发出请求，这样的话它就符合最小化风险的标准，因为其请求是在人们正常浏览网页期间触发的（Narayanan and Zevenbergen 2015）。

当他们决定进行具有未知风险的研究时，第二个重要思想是效果分析，它允许研究人员计算他们所需要的样本大小，从而可靠地检测对给定大小的样本的影响（Cohen 1988）。如果你的研究可能使参与者面临风险，即使是最小的风险，那么根据有利化原则，你也应该为实现研究目标而设置最小的风险（回到第4章中的减少参与者原则）。尽管一些研究人员倾向于让他们的研究规模尽可能大，但研究伦理建议研究规模应该尽可能小。效果分析当然不是新功能，但它在模拟时代的使用方式与今天有着重要的区别。在模拟时代，研究人员通过进行效果分析，确保他们的研究规模不是太小（即效能不足）。然而，现在的研究人员应该利用效果分析确保他们的实验规模不会过大（即效能过剩）。

最低风险标准和效果分析可以帮助你衡量和设计研究，但是它们无法提供任何有关参与者如何看待你的研究以及他们参与研究会遇到什么风险之类的新信息。处理不确定性的另一种方法就是搜集更多的信息，即进行道德反应调查与阶段性测试。

在道德反应调查中，研究人员会对提议的研究项目进行简要描述，然后提出两个问题：

- **问题1：**“如果你关心的人是这个实验的候选参与者，你是否希望他成为参与者？”回答是、无所谓、否；

- **问题2：**“你认为应该允许研究人员继续这个实验吗？”回答是、是（但是要注意）、不确定、否。

在每个问题被回答之后，回答者都可以解释他们的答案。最后，可能成为参与者的人或可能从微任务劳动力市场（如机器人MTurk）被招募的受访者也会回答一些基本的人口统计学问题。

道德反应调查有三个特点，我认为特别具有吸引力。首先，它们在研究之前就已经发生，因此可以在研究开始之前预防问题的产生（这与监测不良反应的方法相反）。其次，道德反应调查的受访者通常不是研究人员，因此这有助于研究人员从公众的角度看待他们的研究。最后，道德反应调查使研究人员能够提出多个版本的研究项目，以评价不同版本对同一项目的伦理平衡。尽管如此，道德反应调查的一个局限性是，在调查结果给出的不同研究设计之间，如何做出决定，它并不明确。但是，忽略这种局限性，道德反应调查不失为是有帮助的；事实上，舍希特尔（Schechter）和布拉沃·利洛（Bravo-Lillo）就放弃了一项计划中的研究，以回应参与者在道德反应调查中提出的问题。

虽然道德反应调查有助于评估对计划研究的反应，但它们无法衡量不良事件的可能性或严重程度。医学研究人员处理高风险环境下不确定性的一种方法是进行阶段性实验，这种方法可能对某些社会研究有帮助。当测试新药的有效性时，研究人员不会立即跳至大规模的随机对照实验的阶段。相反，他们首先进行两种类型的研究。最初，在I期试验中，研究人员特别关注寻找安全剂量，这阶段研究仅涉及少数人。一旦确定了安全剂量，II期试验就会评估该药物的疗效，即评估其在最佳情况下的有效性（Singal, Higgins, and Waljee 2014）。只有在I期与II期试验完成后，新的药物才被允许投入大规模的随机对照实验中。虽然用于开发新药的分阶段实验的确切结构可能不适合用来进行社会研究，但当面临不确定性时，研究人员可以开展针对安全性和有效性的小规模研究。譬如，在“Encore”项目中，你可以想象研究人员从来自更讲究法治的国家的参与者开始研究。

总之，这4种方法，即最低风险标准、效果分析、道德反应调查以及分阶段实验，即使在面对不确定性的情况下，都可以帮助你以合理的方式进行研究。不确定性并不一定导致无所作为。

6.7 实用技巧

除了高尚的道德原则以外，研究道德伦理还存在实际操作问题。

除了本章描述的道德原则与道德框架之外，我还想根据我在数字时代推动、审查和讨论的社会研究中的个人经验，提供三条实用技巧：机构审查委员会是底线，不是上线；换位思考；将研究伦理视作连续的而非离散的过程。

6.7.1 机构审查委员会是底线，不是上线

一方面，许多研究人员似乎与机构审查委员会持相反的观点；另一方面，他们认为机构审查委员会就是装模作样的官僚机构。然而，与此同时，他们也认为它是伦理问题决策的最终仲裁者。也就是说，大多数研究人员似乎认为一旦机构审查委员会通过了审查，那这样做就没问题。如果我们承认机构审查委员会目前存在的非常真实的局限性，并且很多人也这样认为（Schrag 2010, 2011; Hoonaard 2011; Klitzman 2015; King and Sands 2015; Schneider 2015），那么我们作为研究人员必须为研究道德承担额外的责任。机构审查委员会是底线，不是上线，这个想法有两个主要含义。

首先，机构审查委员会是底线意味着，如果你在需要机构审查委员会监管的部门工作，那么你应该遵循这些规定。这似乎是显然的，但是我注意到有些人似乎希望能够避开机构审查委员会。事实上，如果你在伦理上不确定的领域工作，那么机构审查委员会可以成为一个强大的盟友。如果你遵循他们的原则，即使在你的研究出了问题时，

他们也应该支持你（King and Sands 2015）。如果你不遵守规则，可能就要在非常困难的情况下自行解决。

其次，机构审查委员会不是上线意味着，只填写表格并遵守规则是不够的。在许多情况下，你作为研究人员，应该是最了解如何遵守道德规范的人。最终，作为研究人员，道德责任在于你，你的名字会被写在文献上。

确保你将机构审查委员会作为底线而不是上线的一种办法是在论文中加入道德附录。事实上，你可以在研究开始之前就起草道德附录，以便强迫自己考虑如何向同事和公众解释你的工作。如果你在起草道德附录时发现自己感到不适，那么你的研究可能未达到适当的伦理平衡。除了帮助你判断工作外，公布道德附录还有助于研究界讨论伦理问题，并根据真实实证研究中的实例建立适当的规范。表6.3罗列了我认为对伦理研究有良好讨论价值的实证研究论文。我并不同意这些论文作者在讨论中提出的所有声明，但是他们都是卡特（Carter 1996）定义下的完整性研究人员的例子：在每个例子中，（1）研究人员都决定出他们认为是对的和错的事情；（2）他们根据自己的决定采取行动，即使是在个人成本方面；（3）他们公开表示的行为是基于对情境的道德分析的。

表6.3 关于伦理引发有趣讨论的论文

研究人员	研究项目
van de Rijt et al. (2014)	未经同意的实地调查 避免情境伤害
Paluck and Green (2009)	发展中国家的实地调查 敏感话题的研究 复杂的同意议题 对可能造成伤害的补救
Burnett and Feamster (2015)	未经同意的研究 在风险难以评估的情况下进行风险收益权衡
Chaabane et al. (2014)	社会影响的研究 使用泄露的数据文件
Jakobsson and Ratkiewicz (2006)	未经同意的实地调查
Soeller et al. (2016)	违反服务条款

6.7.2 换位思考

研究人员通常非常关注其工作的科学目标，他们只能通过这个角度看到世界。这样的短视可能会造成伦理上的错误判断。因此，当你思考你的研究时，试想一下，你的参与者、利益相关者甚至是记者会对研究做出什么样的反应。这种换位思考与你试想在他们位置的感受是不同的。相反，你试图想象其他人会如何感受，这个过程可能引发同理心（或者叫换位思考）（Batson, Early, and Salvarani 1997）。从这些不同的角度思考你的工作，可以帮助你面对问题并让你的工作具备更好的道德平衡。

此外，当从别人的角度想象你的工作时，你应该期望他们可能注意到某些具体的、糟糕的情况。譬如，为了回应情绪感染项目，一些

批评家专注于那些可能造成自杀的可能性，这是一种低概率但是很极端的糟糕情况。一旦人们的情绪受到刺激而关注最坏的情况，他们可能会完全放弃这种糟糕情况之外的可能性（Sunstein 2002）。然而，人们可能在情绪上做出反应的事实并不意味着你应该将他们视为不知情、非理性或是愚蠢的。我们都应该谦逊地意识到，我们之中没有一个人拥有完美的道德观。

6.7.3 将研究伦理视作连续的而非离散的过程

数字时代社会研究的伦理争议经常是二元（对与错）的。譬如，情绪感染项目是一个要么道德要么不道德的项目。这种二元思维使讨论变得极端，阻碍了开发共享规范的努力，使思想懒惰。研究被打上了“道德”的标签，这使研究人员免除了更加道德地行事的责任。我所见过的涉及研究伦理的最有成效的对话超越了这种二元思维，成为关于研究伦理的一个连续的概念。

研究伦理二元概念的一个主要实际问题是它会使讨论变得偏激。把情绪感染项目称为“不道德的”，会以一种无益的方式将它与真正的暴行混为一谈。相反，更具体地讨论研究中遇到的问题会更有帮助并且更恰当。摆脱二元思维和偏激的语言并不代表我们要用模棱两可的语言隐藏不道德的行为。相反，我认为，连续的道德概念将带来更加谨慎和精确的语言。此外，研究伦理的连续概念可以厘清这样一个概念：每个人，甚至那些正在从事“道德的”工作的研究人员，都应该努力在其工作中创造更好的道德平衡。

迈向持续思考的最终益处在于，它鼓励了谦逊的智慧，而谦逊的智慧在遇到困难的道德挑战时是有益的。数字时代的研究伦理问题是复杂的，任何一个人都不应该对自己判断正确行为的能力过于自信。

6.8 结论

数字时代的社会研究引发了新的伦理问题，但这些问题并非不可解决。作为一个社群，如果我们可以制定由研究人员和公众支持的共同道德准则和标准，那么就能以对社会负责任和有益的方式利用数字时代的能力。本章表达了我试图将我们推向这个方向的想法，并且我认为，关键是研究人员应该采取基于原则的思维方式，同时继续遵守适当的规则。

在6.2节中，我描述了引起道德争议的三个数字时代研究项目。接下来，在6.3节中，我描述了数字时代社会研究中伦理不确定性的根本原因：研究人员在未经参与者同意甚至在其没有意识到的情况下对人们进行观察和实验的能力在迅速增强。这些能力的变化速度远超我们的规范、规则和法律的修订速度。再者，在6.4节中，我描述了四个可以指导你思想的既有原则：对人的尊重原则、有利化原则、公正原则和对法律和公共利益的尊重原则。在6.5节中，我归纳了两种广泛的道德框架，结果主义与义务论，这可以帮助你解决你可能面临的最深刻的挑战之一：何时你适合采取伦理上有问题的手段来达到符合道德标准的目的。这些原则和道德框架将使你超越现有法规所允许的范围去看问题，并提高你向其他研究人员和公众表达你的判断的能力。

基于这样的背景，在6.6节中，我讨论了数字时代研究人员面临的四个特别的挑战：知情同意（6.6.1小节）、理解与管理信息化风险（6.6.2小节）、隐私权（6.6.3小节），以及面对不确定性做出道德决策（6.6.4小节）。最后，在6.7节中，我归纳了三项实用技巧，以应对在不稳定的道德领域工作的情况。

在整体范围方面，本章集中于从独立研究人员的角度寻求可概括性的知识。因此，它产生了关于改进研究伦理监督体系的重要问题、关于管理公司搜集和使用数据的问题，以及对政府大规模监测的质疑。这些问题显然是复杂和困难的，但我希望研究伦理的一些观点对这些其他背景下的研究有所裨益。

历史附录

该历史附录简要回顾了美国研究伦理方面的历史。

任何关于研究伦理的讨论都需要承认，在过去，某些研究人员以科学的名义做了可怕的事情。这其中最糟糕的就是塔斯基吉梅毒实验（表6.4）。1932年，来自美国公共卫生局的研究人员在一项研究中招募了约400名感染梅毒的黑人男性，以监测该疾病的影响。这些男性来自亚拉巴马州的塔斯基吉。从一开始，这项研究就是非治疗性的，它的目的仅仅是记录黑人男性的疾病史。参与者被隐瞒了研究的性质，他们被告知这是一项败血症研究，并且被提供了虚假的和无效的治疗，而梅毒本身是一种致命疾病。随着研究的深入，人们开发出了安全有效的梅毒治疗方法，但该实验的研究人员积极干预以防止参与者在其他地方接受治疗。例如，在第二次世界大战期间，该研究小组在研究中确保所有人在研究期间缓服兵役，以防止这些男性进入部队时获得治疗。研究人员40年中持续欺骗参与者并拒绝治疗他们。

表6.4 塔斯基吉梅毒实验的部分时间线

时间	事件
1932 年	约 400 名感染梅毒的男性被招募至研究中，他们并未被告知研究的真实目的
1937—1938 年	美国公共卫生局向该地区派遣了移动治疗单位，但拒绝治疗研究中的男性
1942—1943 年	为防止这些受试男性在研究中接受其他地方的治疗，公共卫生局在第二次世界大战期间介入，防止他们服兵役
20 世纪 50 年代	青霉素开始成为治疗梅毒的广泛有效的措施，但这些受试男性并未接受治疗（Brandt 1978）
1969 年	美国公共卫生局召开对该研究的伦理审查，审查小组建议继续进行该实验
1972 年	美国公共卫生局前雇员彼得·巴克斯顿（Peter Buxtun）透露了该实验，并在媒体上公布这一消息
1972 年	美国参议院召开有关人类实验的听证会，包括塔斯基吉梅毒实验
1973 年	美国政府停止该实验并责令对幸存者进行治疗
1997 年	美国总统比尔·克林顿公开塔斯基吉梅毒实验并进行官方道歉

塔斯基吉梅毒实验是在当时美国南部地区常见的种族主义和极端不平等背景下进行的。但是，在40年的历史中，这项研究涉及数十名黑人受试者和白人研究人员。除了直接参与的研究人员以外，还有很多人肯定在已发表的医学文献中阅读过相关的15篇研究报告中的某一篇（Heller 1972）。在20世纪60年代中期，也就是研究开始大约30年后，一位名叫彼得·巴克斯顿的美国公共卫生局雇员开始在其内部推动结束这一研究，他认为这种研究在伦理上令人无法接受。作为对巴克斯顿的回应，美国公共卫生局在1969年召集了一个小组，对该研究进行了完整的伦理审查。令人震惊的是，伦理审查小组认为研究人员应该继续拒绝给受感染的男性提供治疗。在评议的过程中，专家组的一位成员甚至表示：“你永远不会再有这样研究的机会，好好利用它

吧。”（Brandt 1978）。这个绝大多数由博士组成的白人专家组认为应该获取某种形式的知情同意。但是该专家组也认为，由于受试者的年龄和低教育程度问题，他们自己无法提供知情同意。因此，专家组建议研究人员从当地医疗官员处获得“代理人知情同意”。所以，经过全面的伦理审查，继续治疗的建议被驳回。最终，巴克斯顿将这件事告诉了一位记者。1972年，简·海勒（Jean Heller）撰写了一系列报道文章，向全世界揭示了这项研究。在广泛的公众愤怒情绪之下，这项研究才最终结束，那些幸存下来的男性才得到治疗。

该研究的受害者并不只是这些男性，还包括他们的家庭：至少22名妻子、17名儿女以及2名孙子，他们均可能由于没有受到治疗而感染梅毒（Yoon 1997）。更进一步，这项研究造成的伤害在其结束后也持续了很长时间。该研究在法理上减弱了非裔美国人对医学界的信任，这种信任的崩塌可能导致非裔美国人拒绝医疗护理而损害他们的健康（Alsan and Wanamaker 2016）。此外，缺乏信任阻碍了在20世纪80年代和90年代治疗艾滋病的努力（Jones 1993，第14章）。

尽管我们今天很难想象会有如此可怕的研究发生，但我认为塔斯基吉梅毒实验对于在数字时代进行社会研究的人有三个重要的经验教训。首先，它提醒我们，有些研究根本不应该发生。其次，它向我们表明，有些研究可能不只对参与者造成伤害，还会在研究结束后对他们的家庭以及整个社群造成长期伤害。最后，它告诉我们，某些研究人员也可能做出很可怕的道德决定。事实上，我认为今天研究人员应该感到一些恐惧，因为参与这项研究的很多人在如此长的时间内做出并坚持了如此糟糕的决定。并且，不幸的是，塔斯基吉的例子并不是唯一的，那个时代还存在着许多在社会和医疗研究中相似的有问题的事例（Katz, Capron, and Glass 1972; Emanuel et al. 2008）。

1974年，为了回应塔斯基吉梅毒实验及其研究人员的伦理过失，美国国会成立了生物医学及行为研究人体受试者保护全国委员会

（National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research），并负责制定涉及人类受试者研究的伦理准则。在贝尔蒙会议中心召开会议的4年后，该小组编写了《贝尔蒙报告》，该报告对生物伦理学和日常研究实践均产生了巨大影响。

《贝尔蒙报告》由三部分组成。第一部分，关于实践与研究之间的界限，该报告阐明了其权限范围。特别是，它主张区分获取一般化知识的研究与包括日常措施和行为在内的实践。此外，第一部分讲到《贝尔蒙报告》的道德原则仅适用于研究。有人认为，这种研究和实践之间的区分是《贝尔蒙报告》不适合数字时代社会研究的一个原因（Metcalf and Crawford 2016; boyd 2016）。

《贝尔蒙报告》的第二部分和第三部分提出了三个道德原则，即对人的尊重原则、有利化原则和公正原则，并描述了这些原则要如何应用于研究实践中。这些原则的细节我已经在本章的正文部分讲述了。

《贝尔蒙报告》设定了广泛的目标，但它不是一个可以轻松用于监管日常活动的文件。因此，美国政府制定了一套俗称为《通则》的法规（Porter and Koski 2008）。这些规定描述了审查、批准和监督研究的过程由机构审查委员会负责执行。要理解《贝尔蒙报告》和《通则》之间的区别，请考虑各方如何讨论知情同意：《贝尔蒙报告》描述了知情同意的哲学原因和代表真正意义上知情同意的广泛特征，而《通则》列出了知情同意的8个必要条件和6个可选要素。根据法律，《通则》管辖几乎所有接受美国政府资助的研究项目。此外，许多从美国政府获得资助的机构通常将《通则》应用于该机构进行的所有研究，而不考虑资金来源。但《通则》并不自动适用于没有从美国政府获得研究经费的公司。

我认为几乎所有研究人员都尊重《贝尔蒙报告》所表达的伦理研究的广泛目标，但《通则》以及与机构审查委员会合作的过程普遍令人感到烦恼（Schrag 2010, 2011; Hoonaard 2011; Klitzman 2015; King and Sands 2015; Schneider 2015）。需要搞清楚的是，那些批评机构审查委员会的人并不反对道德规范。相反，他们认为目前的体系没有达到适当的平衡，或者可以通过其他方法更好地实现其目标。无论如何，我将会把机构审查委员会视作给定条件。如果你需要遵守机构审查委员会的规则，那么你就应该这样做。尽管如此，我仍旧鼓励你在考虑研究伦理时也采取基于原则的方法。

这一背景非常简要地总结了我们如何在美国遵守机构审查委员会基于规则的制度。当我们今天考虑《贝尔蒙报告》和《通则》时，应该记住它们是在不同的时代被创造的，并且对当时产生的问题，它们做出了相当明确的回应，特别是对“二战”期间和之后的医学伦理学做出了回应（Beauchamp 2011）。

除了医学和行为科学家为创造道德规范所做的努力之外，计算机科学家也做出了一些规模较小且知名度不大的努力。事实上，第一批关注数字时代研究所带来的伦理挑战的研究人员并不是社会科学家，而是计算机科学家，特别是在计算机安全领域的研究人员。在20世纪90年代和21世纪初，计算机安全研究人员进行了一系列有道德的研究，这些研究涉及接管僵尸网络和侵入成千上万台弱加密的计算机（Bailey, Dittrich, and Kenneally 2013; Dittrich, Carpenter, and Karir 2015）。针对这些研究，美国政府部门，特别是美国国土安全部设立了一个蓝带委员会，为涉及信息和通信技术的研究撰写指导性道德框架。其成果就是《门罗报告》（Dittrich, Kenneally, and others 2011）。尽管计算机安全研究人员的担忧与社会研究人员的担忧不尽相同，但《门罗报告》为社会研究人员提供了三个重要的指导。

首先，《门罗报告》再次重申了《贝尔蒙报告》中涉及的三项基本原则，即对人的尊重原则、有利化原则以及公正原则，并附带了一项新的原则：对法律和公共利益的尊重原则。我在本章正文中描述了第四项原则以及它如何被应用于社会研究（6.4.4小节）。

其次，《门罗报告》呼吁研究人员超越《贝尔蒙报告》中“涉及人体科学的研究”的狭隘定义，转而采用“具有潜在人身伤害的研究”这一更普遍的概念。“Encore”项目很好地说明了《贝尔蒙报告》定义范围的局限性。普林斯顿大学和乔治亚理工学院的机构审查委员会裁定“Encore”项目不是“涉及人体科学的研究”，因此不受《通则》的监管。然而，“Encore”项目显然具有人身伤害的潜力；在最极端的情况下，“Encore”项目可能会导致无辜的人被专制政府监禁。基于原则的方法意味着，即使机构审查委员会同意，研究人员也不应该隐藏在狭隘的法律定义之后。相反，他们应该采用“具有潜在人身伤害的研究”这一更普遍的概念。

第三，《门罗报告》呼吁研究人员扩大在应用《贝尔蒙报告》原则时考虑的相关利益方。随着研究已经从单独的生活领域转移到更加深入日常活动的领域，伦理考虑的范围必须扩展到特定参与者之外，而且应该包括非参与者和研究发生的环境。换言之，《门罗报告》呼吁研究人员扩大他们的道德领域，而不仅仅考虑他们的参与者。

本历史附录提供了对社会科学、医疗科学以及计算机科学研究伦理的简要回顾。有关医疗科学研究伦理的处理方式，请参见伊曼努尔等人（Emanuel et al. 2008）或比彻姆和奇尔德雷斯（Beauchamp and Childress 2012）的长篇著作。

